

ELEVENTH EDITION

economics

**David Begg, Gianluigi Verna,ca,
Stanley Fischer, Rüdiger Dornbusch**

ELEVENTH EDITION

economics

**David Begg, Gianluigi Vernasca,
Stanley Fischer, Rüdiger
Dornbusch**



London Boston Burr Ridge, IL Dubuque, IA Madison, WI New
York San Francisco St. Louis Bangkok Bogotá Caracas Kuala
Lumpur Lisbon Madrid Mexico City Milan Montreal New Delhi
Santiago Seoul Singapore Sydney Taipei Toronto

Economics 11th Edition

David Begg, Gianluigi Vernasca, Stanley Fischer and Rudiger
Dornbusch
ISBN-13 9780077154516
ISBN-10 0077154517



Published by McGraw-Hill Education
Shoppenhangers Road
Maidenhead
Berkshire
SL6 2QL
Telephone: 44 (0) 1628 502 500
Fax: 44 (0) 1628 770 224
Website: www.mcgraw-hill.co.uk

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

The Library of Congress data for this book has been applied for from the
Library of Congress

Executive Editor: Natalie Jacobs
Commissioning Editor: Kiera Jamison
Development Editor: Alexander Krause
Head of Marketing: Vanessa Boddington
Marketing Manager: Geeta Kumar
Head of Production: Beverley Shields

Text design by Hard Lines
Cover design by Adam Renvoize
Printed and bound in Singapore by Markono

Published by McGraw-Hill Education (UK) Limited an imprint of
McGraw-Hill Education, 2 Penn Plaza New York, NY 10121. Copyright
© 2014 by McGraw-Hill Education (UK) Limited. All rights reserved.
No part of this publication may be reproduced or distributed in any form
or by any means, or stored in a database or retrieval system, without the

prior written consent of McGraw-Hill Education (UK) including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Fictitious names of companies, products, people, characters and/or data that may be used herein (in case studies or examples) are not intended to represent any real individual, company, product or event.

ISBN-13 9780077154516

ISBN-10 0077154517

© 2014. Exclusive rights by McGraw-Hill Education for manufacture and export. This book cannot be re-exported from the country to which it is sold by McGraw-Hill Education.

Dedication

For Honora, Mary and Robin – DB
To my family and to my beloved Vitalba – GV

Brief table of contents

- Preface
- Guided Tour
- Boxed Material
- Online Resources
- Connect
- Acknowledgements
- About the Authors

PART ONE Introduction

- 1** Economics and the economy
- 2** Tools of economic analysis
- 3** Demand, supply and the market

PART TWO Positive microeconomics

- 4** Elasticities of demand and supply
- 5** Consumer choice and demand decisions
- 6** Introducing supply decisions
- 7** Costs and supply
- 8** Perfect competition and pure monopoly
- 9** Market structure and imperfect competition
- 10** The labour market
- 11** Factor markets and income distribution
- 12** Risk and information

PART THREE Welfare economics

- 13** Welfare economics
- 14** Government spending and revenue

PART FOUR Macroeconomics

- 15** Introduction to macroeconomics

- 16** Output and aggregate demand
- 17** Fiscal policy and foreign trade
- 18** Money and banking
- 19** Interest rates and monetary transmission
- 20** Monetary and fiscal policy
- 21** Aggregate supply, prices and adjustment to shocks
- 22** Inflation, expectations and credibility
- 23** Unemployment
- 24** Exchange rates and the balance of payments
- 25** Open economy macroeconomics
- 26** Exchange rate regimes
- 27** Business cycles
- 28** Supply-side economics and economic growth

PART FIVE The world economy

- 29** International trade

Appendix: Answers to activity and maths questions

Glossary

Index

Detailed table of contents

Preface
Guided Tour
Boxed Material
Online Resources
Connect
Acknowledgements
About the Authors

PART ONE Introduction

1 Economics and the economy

- 1.1 How economists think about choices
 - 1.2 Economic issues
 - 1.3 Scarcity and the competing use of resources
 - 1.4 The role of the market
 - 1.5 Positive and normative
 - 1.6 Micro and macro
- Summary
Review questions

2 Tools of economic analysis

- 2.1 Economic data
- 2.2 Index numbers
- 2.3 Nominal and real variables
- 2.4 Measuring changes in economic variables
- 2.5 Economic models
- 2.6 Models and data
- 2.7 Diagrams, lines and equations
- 2.8 Another look at ‘other things equal’
- 2.9 Theories and evidence
- 2.10 Some popular criticisms of economics and economists

Summary

Review questions

3 Demand, supply and the market

- 3.1 The market
- 3.2 Demand, supply and equilibrium
- 3.3 Demand and supply curves
- 3.4 Behind the demand curve
- 3.5 Shifts in the demand curve
- 3.6 Behind the supply curve
- 3.7 Shifts in the supply curve
- 3.8 Consumer and producer surplus
- 3.9 Free markets and price controls
- 3.10 What, how and for whom

Summary

Review questions

PART TWO Positive microeconomics

4 Elasticities of demand and supply

- 4.1 The price responsiveness of demand
- 4.2 Price, quantity demanded and total expenditure
- 4.3 Further applications of the price elasticity of demand
- 4.4 Short run and long run
- 4.5 The cross-price elasticity of demand
- 4.6 The effect of income on demand
- 4.7 Inflation and demand
- 4.8 Elasticity of supply
- 4.9 Taxation and elasticity: who really pays the tax?

Summary

Review questions

5 Consumer choice and demand decisions

- 5.1 Demand by a single consumer
- 5.2 Adjustment to income changes
- 5.3 Adjustment to price changes
- 5.4 The market demand curve
- 5.5 Complements and substitutes

5.6 Transfers in kind

Summary

Review questions

Appendix: Consumer choice with measurable utility

6 Introducing supply decisions

6.1 Business organization

6.2 A firm's accounts

6.3 Firms and profit maximization

6.4 The firm's supply decision

6.5 Marginal cost and marginal revenue

6.6 Marginal cost and marginal revenue curves

Summary

Review questions

7 Costs and supply

7.1 Inputs and output: the production function

7.2 Production in the short run: diminishing marginal returns

7.3 Short-run costs

7.4 A firm's output decision in the short run

7.5 Production in the long run

7.6 Long-run total, marginal and average costs

7.7 Returns to scale

7.8 The firm's long-run output decision

7.9 The relationship between short-run and long-run average costs

Summary

Review questions

Appendix: Isoquants and the choice of production technique in the long run

8 Perfect competition and pure monopoly

8.1 Perfect competition

8.2 A perfectly competitive firm's supply decision

8.3 Industry supply curves

8.4 Comparative statics for a competitive industry

8.5 Pure monopoly: the opposite limiting case

8.6 Profit-maximizing output for a monopolist

8.7 Output and price under monopoly and competition

8.8 A monopoly has no supply curve

8.9 Monopoly and technical change

8.10 Natural monopoly

Summary

Review questions

9 Market structure and imperfect competition

9.1 Why market structures differ

9.2 Monopolistic competition

9.3 Oligopoly and interdependence

9.4 Game theory and interdependent decisions

9.5 Reaction functions

9.6 Entry and potential competition

9.7 Strategic entry deterrence

9.8 Summing up

Summary

Review questions

10 The labour market

10.1 The firm's demand for factors in the long run

10.2 The firm's demand for labour in the short run

10.3 The industry demand curve for labour

10.4 The supply of labour

10.5 Industry labour market equilibrium

10.6 Transfer earnings and economic rents

10.7 Do labour markets clear?

10.8 Wage discrimination

Summary

Review questions

11 Factor markets and income distribution

11.1 Physical capital

11.2 Rentals, interest rates and asset prices

11.3 Saving, investment and the real interest rate

11.4 The demand for capital services

11.5 The supply of capital services

11.6 Equilibrium and adjustment in the market for capital services

11.7 The price of capital assets

11.8 Land and rents

11.9 Income distribution in the UK

Summary

Review questions

12 Risk and information

12.1 Individual attitudes to risk

12.2 Insurance and risk

12.3 Asymmetric information

12.4 Uncertainty and asset returns

12.5 Portfolio selection

12.6 Efficient asset markets

12.7 More on risk

Summary

Review questions

PART THREE Welfare economics

13 Welfare economics

13.1 Equity and efficiency

13.2 Perfect competition and Pareto efficiency

13.3 Distortions and the second-best

13.4 Market failure

13.5 Externalities

13.6 Environmental issues and the economics of climate change

13.7 Other missing markets: time and risk

13.8 Quality, health and safety

Summary

Review questions

14 Government spending and revenue

14.1 Taxation and government spending

14.2 The government in the market economy

14.3 The principles of taxation

14.4 Taxation and supply-side economics

14.5 Local government

14.6 Economic sovereignty

14.7 Political economy: how governments decide

Summary

Review questions

PART FOUR Macroeconomics

15 Introduction to macroeconomics

- 15.1** The scope of macroeconomics: the big issues
 - 15.2** Some facts and economic history
 - 15.3** The circular flow
 - 15.4** National income accounting
 - 15.5** What GDP measures
 - 15.6** International comparisons
- Summary
- Review questions

16 Output and aggregate demand

- 16.1** Components of aggregate demand
 - 16.2** Aggregate demand
 - 16.3** Equilibrium output
 - 16.4** Planned saving equals planned investment
 - 16.5** A fall in aggregate demand
 - 16.6** The multiplier
 - 16.7** The paradox of thrift
 - 16.8** The role of confidence
- Summary
- Review questions

17 Fiscal policy and foreign trade

- 17.1** The scope of government activity
- 17.2** Government and aggregate demand
- 17.3** The government budget
- 17.4** Deficits and the fiscal stance
- 17.5** Automatic stabilizers and discretionary fiscal policy
- 17.6** The national debt and the deficit
- 17.7** Foreign trade and income determination

Summary

Review questions

18 Money and banking

- 18.1** Money and its functions
- 18.2** Modern banking

- 18.3** How banks create money
 - 18.4** The traditional theory of money supply
 - 18.5** The demand for money
 - 18.6** Financial crises
- Summary
- Review questions

19 Interest rates and monetary transmission

- 19.1** The Bank of England
 - 19.2** Traditional means of monetary control
 - 19.3** Lender of last resort
 - 19.4** Equilibrium in financial markets
 - 19.5** Monetary control
 - 19.6** Targets and instruments of monetary policy
 - 19.7** The transmission mechanism
- Summary
- Review questions

20 Monetary and fiscal policy

- 20.1** Monetary policy
 - 20.2** The *IS–LM* model
 - 20.3** The *IS–LM* model in action
 - 20.4** Shocks to money demand
 - 20.5** The policy mix
 - 20.6** The effect of future taxes
 - 20.7** Demand management revisited
- Summary
- Review questions

21 Aggregate supply, prices and adjustment to shocks

- 21.1** Inflation and aggregate demand
- 21.2** Aggregate supply
- 21.3** Equilibrium inflation
- 21.4** The labour market and wage behaviour
- 21.5** Short-run aggregate supply
- 21.6** The adjustment process
- 21.7** Sluggish adjustment to shocks
- 21.8** Trade-offs in monetary objectives

Summary

Review questions

22 Inflation, expectations and credibility

- 22.1 Money and inflation
- 22.2 Inflation and interest rates
- 22.3 Inflation, money and deficits
- 22.4 Inflation, unemployment and output
- 22.5 The costs of inflation
- 22.6 Controlling inflation
- 22.7 The Monetary Policy Committee

Summary

Review questions

23 Unemployment

- 23.1 The labour market
- 23.2 Analysing unemployment
- 23.3 Explaining changes in unemployment
- 23.4 Cyclical fluctuations in unemployment
- 23.5 The cost of unemployment

Summary

Review questions

24 Exchange rates and the balance of payments

- 24.1 The foreign exchange market
- 24.2 Exchange rate regimes
- 24.3 The balance of payments
- 24.4 The real exchange rate
- 24.5 Determinants of the current account
- 24.6 The financial account
- 24.7 Internal and external balance
- 24.8 The long-run equilibrium real exchange rate

Summary

Review questions

25 Open economy macroeconomics

- 25.1 Fixed exchange rates
- 25.2 Macroeconomic policy under fixed exchange rates
- 25.3 Devaluation
- 25.4 Floating exchange rates

- 25.5** Monetary and fiscal policy under floating exchange rates
Summary
Review questions

26 Exchange rate regimes

- 26.1** The gold standard
26.2 An adjustable peg
26.3 Floating exchange rates
26.4 Speculative attacks on fixed exchange rates
26.5 Antecedents of the Eurozone
26.6 The economics of the euro

Summary
Review questions

27 Business cycles

- 27.1** Trend and cycle: statistics or economics
27.2 Theories of the business cycle
27.3 Real business cycles
27.4 Supply-side effects of the financial crash
27.5 An international business cycle?
27.6 Summing up: schools of macroeconomic thinking

Summary
Review questions

28 Supply-side economics and economic growth

- 28.1** Supply-side economics
28.2 Economic growth: preliminary remarks
28.3 Growth: an overview
28.4 Technical knowledge
28.5 Growth and accumulation
28.6 Growth through technical progress
28.7 Growth in the OECD
28.8 Endogenous growth
28.9 The costs of growth

Summary
Review questions

PART FIVE The world economy

29 International trade

- 29.1** Trade patterns
- 29.2** Comparative advantage
- 29.3** Intra-industry trade
- 29.4** The economics of tariffs
- 29.5** Good and bad arguments for tariffs
- 29.6** Tariff levels: not so bad?
- 29.7** Other trade policies

Summary

Review questions

Appendix: Answers to activity and maths questions

Glossary

Index

Preface

Economics is much too interesting to be left to professional economists. It affects almost everything we do, not merely at work or at the shops but also in the home and the voting booth. It influences how well we look after our planet, the future we leave for our children, the extent to which we can care for the poor and the disadvantaged, and the resources we have for enjoying ourselves.

These issues are discussed daily, in bars and on buses as well as in cabinet meetings and boardrooms. The formal study of economics is exciting because it introduces a toolkit that allows a better understanding of the problems we face. Everyone knows a smoky engine is a bad sign, but sometimes only a trained mechanic can give the right advice on how to fix it.

This book is designed to teach you the toolkit and give you practice in using it. Nobody carries an enormous toolbox very far. Useful toolkits are small enough to be portable but contain enough proven tools to deal with both routine problems and unforeseen circumstances. With practice, you will be surprised at how much light this analysis can shed on daily living. This book is designed to make economics seem as useful as it really is.

How much do economists disagree?

People often complain that economists never agree about anything. This is simply not true. The media, taxi-drivers and politicians love to talk about topics on which there is disagreement; it would be boring TV if all participants in a panel discussion held identical views. But economics is not a subject on which there is always an argument for everything. There are answers to many questions.

We aim to show where economists agree – on what and for what reason – and why they sometimes disagree.

Economics in the twenty-first century

Our aim is to allow students to understand today's economic environment. This requires mastering the theory and practising its application. Just as the theory of genetics or of information technology is slowly progressing, so the theory of economics continues to make progress, sometimes in dramatic and exciting ways. Sometimes this is prompted by theoretical reasoning; sometimes it is a response to a dramatic new event, such as the banking crash and subsequent financial meltdown around the world.

We believe in introducing students immediately to the latest ideas in economics. If these can be conveyed simply, why force students to use older approaches that work less well? Two recent developments in economics underlie much of what we do. One is the role of information, the other is globalization.

How information is transmitted and manipulated is central to many issues concerning incentives and competition, including the recent boom in e-commerce. Ease of information, coupled with lower transport costs, also explains trends towards globalization, and associated reductions in national sovereignty, especially in smaller countries. Modern economics helps us make sense of our changing world, think about where it may go next, and evaluate choices that we currently face.

Changes to the eleventh edition

After 30 years at the top, we wanted to ensure that the book remains as relevant for the next 30 years as it has been in the past. Those familiar with previous editions will continue to recognize the underlying structure and approach, a window on the latest thinking about our evolving world and the way in which economics can make sense of it.

The eleventh edition has again been thoroughly revised, using feedback from a broad range of actual and potential adopters. A lot has changed in the last three years and it is all reflected in the text. Specific changes to the new edition include:

- A more detailed introduction to economic analysis has been added to Chapter 1.
- A revision of the analysis of consumer choice and demand decisions.
- A more detailed analysis of cost minimization.

- An extended analysis of individuals' labour supply and wage discrimination.
- A comprehensive analysis of the financial crash – its causes, consequences and possible policy responses, from fiscal stimulus to quantitative easing.
- Full updates throughout to include 2012/13 data in graphs and tables.
- Over 160 contemporary case studies, activities, concepts etc., over 80 of which are new, and which illustrate key ideas with relevance to the real world.
- Important new pedagogical features, including topical new case studies, boxes on economic concepts and activity applications, and optional maths boxes for the technically-minded.
- More electronic resources for both students and lecturers.

With all this change, the book's structure is a bit different from the last edition. The main changes are listed below:

- In response to user feedback, the previous Chapter 29 has become Chapter 26, making the three open economy chapters contiguous (24–26), and allowing a summing-up in Chapter 27 (business cycles) and 28 (supply-side and long-run growth).
- This edition also contains a discussion of competing schools of macroeconomic thought, allowing students to enrich their perspective on alternative approaches to macroeconomics. Chapter 16 briefly reviews the history of macroeconomic thought and Chapter 27 takes stock of how different schools of thought interpret output fluctuations and deviations from potential output.
- Chapter 28 contains an expanded discussion of the role of supply-side economics.
- Having absorbed open economy macroeconomics into our core discussion, the role of Part Five (on the world economy) is now confined to examination of international trade issues.

Learning by doing

Few people practise for a driving test just by reading a book. Even when you think you understand how to do a hill start, it takes a lot of practice to master the finer points. In the same way, we give you lots of examples and real-world applications not just to emphasize the relevance of economics but also to help you master it for yourself. We start at square one and take you slowly through the tools of theoretical reasoning and how to apply them. We use algebra and equations sparingly, more often than not in separate boxes so they can be skipped or used depending on how technically-minded you are. The best ideas are simple and robust, and can usually be explained quite easily.

How to study

Don't just read about economics, try to do it! It is easy, but mistaken, to read on cruise control, highlighting the odd sentence and gliding through paragraphs we have worked hard to simplify. Active learning needs to be interactive. When the text says 'clearly', ask yourself 'why' it is clear. See if you can construct the diagram before you look at it. As soon as you don't follow something, go back and read it again. Try to think of other examples to which the theory could be applied. The only way to check you really understand things is to test yourself. There are opportunities to do this in the book through activity and review questions, as well as online through Connect Economics.

To assist you in working through this text, we have developed a number of distinctive study and design features. To familiarize yourself with these features, please turn to the Guided Tour on pages xv–xvii.

Supplementary resources

Economics, eleventh edition offers a comprehensive package of resources for the teaching and learning of economics. The resources offered with the new edition have been developed in response to feedback from current users in order to provide lecturers with a variety of teaching resources for class teaching, lectures and assessment. Students are also offered a range of extra materials to assist in learning, revising and applying the principles of economics.

Connect

Connect Economics is a digital teaching and learning environment that improves performance over a variety of critical outcomes. It gives lecturers and tutors the power to create assignments, tests and quizzes online. Easily accessible grade reports allow you to track your students' progress. Students get feedback on each individual question and immediate grading on both assignments and independent revision questions, which makes it the perfect platform to test your knowledge. Videos, animated graphs, case studies and vignettes allow students to explore complicated graphs and concepts in more detail.

To learn more about Connect, turn to page xxiv or visit connect.mcgraw-hill.com/economics.

Online Learning Centre

An accompanying Online Learning Centre website has been developed to provide an unrivalled package of flexible, high-quality resources for lecturers. To access all of the free Online Learning Centre resources and to find out about enhanced options, simply visit the website at www.mcgraw-hill.co.uk/textbooks/begg.

To learn more about the resources available to lecturers and students online, go to our tour of the resources on page xxiii.

Guided tour

The image shows two screenshots of the Connect platform. The left screenshot displays a 'Review questions' section with a list of 7 numbered questions. The right screenshot shows the 'library' section, listing various chapters and their corresponding assignment counts.

Review questions

- True or False. On a given indifference curve, the marginal rate of substitution is always decreasing.
- Common fallacies. Why are these statements wrong? (a) Since consumers do not know about indifference curves or budget lines, they cannot choose the point on the budget line tangent to the highest possible indifference curve. (b) Inflation must reduce demand since prices are higher and goods are more expensive.
- Suppose there are normal goods but transport is an inferior good. How do the quantities demanded for the two goods change when income increases?
- The own-price elasticity of demand for food is negative. The demand for food is elastic. A higher food price increases spending on food. Higher food prices imply less is spent on all other goods. The quantity demanded of each of these other goods falls. Discuss each statement. Are they all correct?
- Suppose Gleneagles have a given income and like weekend trips to the Highlands, which are a three-hour drive away. (a) If the price of petrol doubles, what is the effect on the demand for trips to the Highlands? Discuss both income and substitution effects. (b) What happens to the demand for Highland hotel rooms?
- Frank's utility function for two goods, X and Y , is given by $U = XY$. Find Frank's indifference curves, when utility is 10, 20 and 30. Plot these indifference curves. How should Frank compare the following two bundles: ($X = 1, Y = 10$) and ($X = 5, Y = 21$)?
- Suppose Frank has an income of £50, the unit price of X is $P_X = £2$ and the unit price of Y is $P_Y = £1$. Write down the budget constraint for Frank. Knowing that the marginal rate of substitution (in absolute value) between X and Y is MRS = 20, find the optimal bundle that Frank should consume. (Check it.)

library

chapter	sections	questions	answers	instruct	work	practice
Ch01-01INT Chapter Name to begin a chapter	1.1	1	1	Not Yet	Not Yet	Not Yet
Ch01-02INT Chapter Name to begin a chapter	1.2	1	1	Not Yet	Not Yet	Not Yet
Ch01-03INT Chapter Name to begin a chapter	1.3	1	1	Not Yet	Not Yet	Not Yet
Ch01-04INT Chapter Name to begin a chapter	1.4	1	1	Not Yet	Not Yet	Not Yet
Ch01-05INT Chapter Name to begin a chapter	1.5	1	1	Not Yet	Not Yet	Not Yet
Ch01-06INT Chapter Name to begin a chapter	1.6	1	1	Not Yet	Not Yet	Not Yet
Ch01-07INT Chapter Name to begin a chapter	1.7	1	1	Not Yet	Not Yet	Not Yet
Ch01-08INT Chapter Name to begin a chapter	1.8	1	1	Not Yet	Not Yet	Not Yet
Ch01-09INT Chapter Name to begin a chapter	1.9	1	1	Not Yet	Not Yet	Not Yet
Ch01-10INT Chapter Name to begin a chapter	1.10	1	1	Not Yet	Not Yet	Not Yet
Ch02-01INT Chapter Name to begin a chapter	2.1	1	1	Not Yet	Not Yet	Not Yet
Ch02-02INT Chapter Name to begin a chapter	2.2	1	1	Not Yet	Not Yet	Not Yet
Ch02-03INT Chapter Name to begin a chapter	2.3	1	1	Not Yet	Not Yet	Not Yet
Ch02-04INT Chapter Name to begin a chapter	2.4	1	1	Not Yet	Not Yet	Not Yet
Ch02-05INT Chapter Name to begin a chapter	2.5	1	1	Not Yet	Not Yet	Not Yet
Ch02-06INT Chapter Name to begin a chapter	2.6	1	1	Not Yet	Not Yet	Not Yet
Ch02-07INT Chapter Name to begin a chapter	2.7	1	1	Not Yet	Not Yet	Not Yet
Ch02-08INT Chapter Name to begin a chapter	2.8	1	1	Not Yet	Not Yet	Not Yet
Ch02-09INT Chapter Name to begin a chapter	2.9	1	1	Not Yet	Not Yet	Not Yet
Ch02-10INT Chapter Name to begin a chapter	2.10	1	1	Not Yet	Not Yet	Not Yet
Ch02-11INT Chapter Name to begin a chapter	2.11	1	1	Not Yet	Not Yet	Not Yet
Ch02-12INT Chapter Name to begin a chapter	2.12	1	1	Not Yet	Not Yet	Not Yet
Ch02-13INT Chapter Name to begin a chapter	2.13	1	1	Not Yet	Not Yet	Not Yet
Ch02-14INT Chapter Name to begin a chapter	2.14	1	1	Not Yet	Not Yet	Not Yet
Ch02-15INT Chapter Name to begin a chapter	2.15	1	1	Not Yet	Not Yet	Not Yet
Ch02-16INT Chapter Name to begin a chapter	2.16	1	1	Not Yet	Not Yet	Not Yet
Ch02-17INT Chapter Name to begin a chapter	2.17	1	1	Not Yet	Not Yet	Not Yet
Ch02-18INT Chapter Name to begin a chapter	2.18	1	1	Not Yet	Not Yet	Not Yet
Ch02-19INT Chapter Name to begin a chapter	2.19	1	1	Not Yet	Not Yet	Not Yet
Ch02-20INT Chapter Name to begin a chapter	2.20	1	1	Not Yet	Not Yet	Not Yet
Ch02-21INT Chapter Name to begin a chapter	2.21	1	1	Not Yet	Not Yet	Not Yet
Ch02-22INT Chapter Name to begin a chapter	2.22	1	1	Not Yet	Not Yet	Not Yet
Ch02-23INT Chapter Name to begin a chapter	2.23	1	1	Not Yet	Not Yet	Not Yet
Ch02-24INT Chapter Name to begin a chapter	2.24	1	1	Not Yet	Not Yet	Not Yet
Ch02-25INT Chapter Name to begin a chapter	2.25	1	1	Not Yet	Not Yet	Not Yet
Ch02-26INT Chapter Name to begin a chapter	2.26	1	1	Not Yet	Not Yet	Not Yet
Ch02-27INT Chapter Name to begin a chapter	2.27	1	1	Not Yet	Not Yet	Not Yet
Ch02-28INT Chapter Name to begin a chapter	2.28	1	1	Not Yet	Not Yet	Not Yet
Ch02-29INT Chapter Name to begin a chapter	2.29	1	1	Not Yet	Not Yet	Not Yet
Ch02-30INT Chapter Name to begin a chapter	2.30	1	1	Not Yet	Not Yet	Not Yet
Ch02-31INT Chapter Name to begin a chapter	2.31	1	1	Not Yet	Not Yet	Not Yet
Ch02-32INT Chapter Name to begin a chapter	2.32	1	1	Not Yet	Not Yet	Not Yet
Ch02-33INT Chapter Name to begin a chapter	2.33	1	1	Not Yet	Not Yet	Not Yet
Ch02-34INT Chapter Name to begin a chapter	2.34	1	1	Not Yet	Not Yet	Not Yet
Ch02-35INT Chapter Name to begin a chapter	2.35	1	1	Not Yet	Not Yet	Not Yet
Ch02-36INT Chapter Name to begin a chapter	2.36	1	1	Not Yet	Not Yet	Not Yet
Ch02-37INT Chapter Name to begin a chapter	2.37	1	1	Not Yet	Not Yet	Not Yet
Ch02-38INT Chapter Name to begin a chapter	2.38	1	1	Not Yet	Not Yet	Not Yet
Ch02-39INT Chapter Name to begin a chapter	2.39	1	1	Not Yet	Not Yet	Not Yet
Ch02-40INT Chapter Name to begin a chapter	2.40	1	1	Not Yet	Not Yet	Not Yet
Ch02-41INT Chapter Name to begin a chapter	2.41	1	1	Not Yet	Not Yet	Not Yet
Ch02-42INT Chapter Name to begin a chapter	2.42	1	1	Not Yet	Not Yet	Not Yet
Ch02-43INT Chapter Name to begin a chapter	2.43	1	1	Not Yet	Not Yet	Not Yet
Ch02-44INT Chapter Name to begin a chapter	2.44	1	1	Not Yet	Not Yet	Not Yet
Ch02-45INT Chapter Name to begin a chapter	2.45	1	1	Not Yet	Not Yet	Not Yet
Ch02-46INT Chapter Name to begin a chapter	2.46	1	1	Not Yet	Not Yet	Not Yet
Ch02-47INT Chapter Name to begin a chapter	2.47	1	1	Not Yet	Not Yet	Not Yet
Ch02-48INT Chapter Name to begin a chapter	2.48	1	1	Not Yet	Not Yet	Not Yet
Ch02-49INT Chapter Name to begin a chapter	2.49	1	1	Not Yet	Not Yet	Not Yet
Ch02-50INT Chapter Name to begin a chapter	2.50	1	1	Not Yet	Not Yet	Not Yet
Ch02-51INT Chapter Name to begin a chapter	2.51	1	1	Not Yet	Not Yet	Not Yet
Ch02-52INT Chapter Name to begin a chapter	2.52	1	1	Not Yet	Not Yet	Not Yet
Ch02-53INT Chapter Name to begin a chapter	2.53	1	1	Not Yet	Not Yet	Not Yet
Ch02-54INT Chapter Name to begin a chapter	2.54	1	1	Not Yet	Not Yet	Not Yet
Ch02-55INT Chapter Name to begin a chapter	2.55	1	1	Not Yet	Not Yet	Not Yet
Ch02-56INT Chapter Name to begin a chapter	2.56	1	1	Not Yet	Not Yet	Not Yet
Ch02-57INT Chapter Name to begin a chapter	2.57	1	1	Not Yet	Not Yet	Not Yet
Ch02-58INT Chapter Name to begin a chapter	2.58	1	1	Not Yet	Not Yet	Not Yet
Ch02-59INT Chapter Name to begin a chapter	2.59	1	1	Not Yet	Not Yet	Not Yet
Ch02-60INT Chapter Name to begin a chapter	2.60	1	1	Not Yet	Not Yet	Not Yet
Ch02-61INT Chapter Name to begin a chapter	2.61	1	1	Not Yet	Not Yet	Not Yet
Ch02-62INT Chapter Name to begin a chapter	2.62	1	1	Not Yet	Not Yet	Not Yet
Ch02-63INT Chapter Name to begin a chapter	2.63	1	1	Not Yet	Not Yet	Not Yet
Ch02-64INT Chapter Name to begin a chapter	2.64	1	1	Not Yet	Not Yet	Not Yet
Ch02-65INT Chapter Name to begin a chapter	2.65	1	1	Not Yet	Not Yet	Not Yet
Ch02-66INT Chapter Name to begin a chapter	2.66	1	1	Not Yet	Not Yet	Not Yet
Ch02-67INT Chapter Name to begin a chapter	2.67	1	1	Not Yet	Not Yet	Not Yet
Ch02-68INT Chapter Name to begin a chapter	2.68	1	1	Not Yet	Not Yet	Not Yet
Ch02-69INT Chapter Name to begin a chapter	2.69	1	1	Not Yet	Not Yet	Not Yet
Ch02-70INT Chapter Name to begin a chapter	2.70	1	1	Not Yet	Not Yet	Not Yet
Ch02-71INT Chapter Name to begin a chapter	2.71	1	1	Not Yet	Not Yet	Not Yet
Ch02-72INT Chapter Name to begin a chapter	2.72	1	1	Not Yet	Not Yet	Not Yet
Ch02-73INT Chapter Name to begin a chapter	2.73	1	1	Not Yet	Not Yet	Not Yet
Ch02-74INT Chapter Name to begin a chapter	2.74	1	1	Not Yet	Not Yet	Not Yet
Ch02-75INT Chapter Name to begin a chapter	2.75	1	1	Not Yet	Not Yet	Not Yet
Ch02-76INT Chapter Name to begin a chapter	2.76	1	1	Not Yet	Not Yet	Not Yet
Ch02-77INT Chapter Name to begin a chapter	2.77	1	1	Not Yet	Not Yet	Not Yet
Ch02-78INT Chapter Name to begin a chapter	2.78	1	1	Not Yet	Not Yet	Not Yet
Ch02-79INT Chapter Name to begin a chapter	2.79	1	1	Not Yet	Not Yet	Not Yet
Ch02-80INT Chapter Name to begin a chapter	2.80	1	1	Not Yet	Not Yet	Not Yet
Ch02-81INT Chapter Name to begin a chapter	2.81	1	1	Not Yet	Not Yet	Not Yet
Ch02-82INT Chapter Name to begin a chapter	2.82	1	1	Not Yet	Not Yet	Not Yet
Ch02-83INT Chapter Name to begin a chapter	2.83	1	1	Not Yet	Not Yet	Not Yet
Ch02-84INT Chapter Name to begin a chapter	2.84	1	1	Not Yet	Not Yet	Not Yet
Ch02-85INT Chapter Name to begin a chapter	2.85	1	1	Not Yet	Not Yet	Not Yet
Ch02-86INT Chapter Name to begin a chapter	2.86	1	1	Not Yet	Not Yet	Not Yet
Ch02-87INT Chapter Name to begin a chapter	2.87	1	1	Not Yet	Not Yet	Not Yet
Ch02-88INT Chapter Name to begin a chapter	2.88	1	1	Not Yet	Not Yet	Not Yet
Ch02-89INT Chapter Name to begin a chapter	2.89	1	1	Not Yet	Not Yet	Not Yet
Ch02-90INT Chapter Name to begin a chapter	2.90	1	1	Not Yet	Not Yet	Not Yet
Ch02-91INT Chapter Name to begin a chapter	2.91	1	1	Not Yet	Not Yet	Not Yet
Ch02-92INT Chapter Name to begin a chapter	2.92	1	1	Not Yet	Not Yet	Not Yet
Ch02-93INT Chapter Name to begin a chapter	2.93	1	1	Not Yet	Not Yet	Not Yet
Ch02-94INT Chapter Name to begin a chapter	2.94	1	1	Not Yet	Not Yet	Not Yet
Ch02-95INT Chapter Name to begin a chapter	2.95	1	1	Not Yet	Not Yet	Not Yet
Ch02-96INT Chapter Name to begin a chapter	2.96	1	1	Not Yet	Not Yet	Not Yet
Ch02-97INT Chapter Name to begin a chapter	2.97	1	1	Not Yet	Not Yet	Not Yet
Ch02-98INT Chapter Name to begin a chapter	2.98	1	1	Not Yet	Not Yet	Not Yet
Ch02-99INT Chapter Name to begin a chapter	2.99	1	1	Not Yet	Not Yet	Not Yet
Ch02-100INT Chapter Name to begin a chapter	2.100	1	1	Not Yet	Not Yet	Not Yet
Ch02-101INT Chapter Name to begin a chapter	2.101	1	1	Not Yet	Not Yet	Not Yet
Ch02-102INT Chapter Name to begin a chapter	2.102	1	1	Not Yet	Not Yet	Not Yet
Ch02-103INT Chapter Name to begin a chapter	2.103	1	1	Not Yet	Not Yet	Not Yet
Ch02-104INT Chapter Name to begin a chapter	2.104	1	1	Not Yet	Not Yet	Not Yet
Ch02-105INT Chapter Name to begin a chapter	2.105	1	1	Not Yet	Not Yet	Not Yet
Ch02-106INT Chapter Name to begin a chapter	2.106	1	1	Not Yet	Not Yet	Not Yet
Ch02-107INT Chapter Name to begin a chapter	2.107	1	1	Not Yet	Not Yet	Not Yet
Ch02-108INT Chapter Name to begin a chapter	2.108	1	1	Not Yet	Not Yet	Not Yet
Ch02-109INT Chapter Name to begin a chapter	2.109	1	1	Not Yet	Not Yet	Not Yet
Ch02-110INT Chapter Name to begin a chapter	2.110	1	1	Not Yet	Not Yet	Not Yet
Ch02-111INT Chapter Name to begin a chapter	2.111	1	1	Not Yet	Not Yet	Not Yet
Ch02-112INT Chapter Name to begin a chapter	2.112	1	1	Not Yet	Not Yet	Not Yet
Ch02-113INT Chapter Name to begin a chapter	2.113	1	1	Not Yet	Not Yet	Not Yet
Ch02-114INT Chapter Name to begin a chapter	2.114	1	1	Not Yet	Not Yet	Not Yet
Ch02-115INT Chapter Name to begin a chapter	2.115	1	1	Not Yet	Not Yet	Not Yet
Ch02-116INT Chapter Name to begin a chapter	2.116	1	1	Not Yet	Not Yet	Not Yet
Ch02-117INT Chapter Name to begin a chapter	2.117	1	1	Not Yet	Not Yet	Not Yet
Ch02-118INT Chapter Name to begin a chapter	2.118	1	1	Not Yet	Not Yet	Not Yet
Ch02-119INT Chapter Name to begin a chapter	2.119	1	1	Not Yet	Not Yet	Not Yet
Ch02-120INT Chapter Name to begin a chapter	2.120	1	1	Not Yet	Not Yet	Not Yet
Ch02-121INT Chapter Name to begin a chapter	2.121	1	1	Not Yet	Not Yet	Not Yet
Ch02-122INT Chapter Name to begin a chapter	2.122	1	1	Not Yet	Not Yet	Not Yet
Ch02-123INT Chapter Name to begin a chapter	2.123	1	1	Not Yet	Not Yet	Not Yet
Ch02-124INT Chapter Name to begin a chapter	2.124	1	1	Not Yet	Not Yet	Not Yet
Ch02-125INT Chapter Name to begin a chapter	2.125	1	1	Not Yet	Not Yet	Not Yet
Ch02-126INT Chapter Name to begin a chapter	2.126	1	1	Not Yet	Not Yet	Not Yet
Ch02-127INT Chapter Name to begin a chapter	2.127	1	1	Not Yet	Not Yet	Not Yet
Ch02-128INT Chapter Name to begin a chapter	2.128	1	1	Not Yet	Not Yet	Not Yet
Ch02-129INT Chapter Name to begin a chapter	2.129	1	1	Not Yet	Not Yet	Not Yet
Ch02-130INT Chapter Name to begin a chapter	2.130	1	1	Not Yet	Not Yet	Not Yet
Ch02-131INT Chapter Name to begin a chapter	2.131	1	1	Not Yet	Not Yet	Not Yet
Ch02-132INT Chapter Name to begin a chapter	2.132	1	1	Not Yet	Not Yet	Not Yet
Ch02-133INT Chapter Name to begin a chapter	2.133	1	1	Not Yet	Not Yet	Not Yet
Ch02-134INT Chapter Name to begin a chapter	2.134	1	1	Not Yet	Not Yet	Not Yet
Ch02-135INT Chapter Name to begin a chapter	2.135	1	1	Not Yet	Not Yet	Not Yet
Ch02-136INT Chapter Name to begin a chapter	2.136	1	1	Not Yet	Not Yet	Not Yet
Ch02-137INT Chapter Name to begin a chapter	2.137	1	1	Not Yet	Not Yet	Not Yet
Ch02-138INT Chapter Name to begin a chapter	2.138	1	1	Not Yet	Not Yet	Not Yet
Ch02-139INT Chapter Name to begin a chapter	2.139	1	1	Not Yet	Not Yet	Not Yet
Ch02-140INT Chapter Name to begin a chapter	2.140	1	1	Not Yet	Not Yet	Not Yet
Ch02-141INT Chapter Name to begin a chapter	2.141	1	1	Not Yet	Not Yet	Not Yet
Ch02-142INT Chapter Name to begin a chapter	2.142	1	1	Not Yet	Not Yet	Not Yet
Ch02-143INT Chapter Name to begin a chapter	2.143	1	1	Not Yet	Not Yet	Not Yet
Ch02-144INT Chapter Name to begin a chapter	2.144	1	1	Not Yet	Not Yet	Not Yet
Ch02-145INT Chapter Name to begin a chapter	2.145	1	1	Not Yet	Not Yet	Not Yet
Ch02-146INT Chapter Name to begin a chapter	2.146	1	1	Not Yet	Not Yet	Not Yet
Ch02-147INT Chapter Name to begin a chapter	2.147	1	1	Not Yet	Not Yet	Not Yet
Ch02-148INT Chapter Name to begin a chapter	2.148	1	1	Not Yet	Not Yet	Not Yet
Ch02-149INT Chapter Name to begin a chapter	2.149	1	1	Not Yet	Not Yet	Not Yet
Ch02-150INT Chapter Name to begin a chapter	2.150	1	1	Not Yet	Not Yet	Not Yet
Ch02-151INT Chapter Name to begin a chapter	2.151	1	1	Not Yet	Not Yet	Not Yet
Ch02-152INT Chapter Name to begin a chapter	2.152	1	1	Not Yet	Not Yet	Not Yet
Ch02-153INT Chapter Name to begin a chapter	2.153	1				

and then recommends specific readings, revision tools and additional practice work.

Chapter 17

Question #13 (of 14) / [View question](#)

Your score for this is 1.00 marks

13. **MCQ**

MCQ policies have inflation targets to ensure the question...
Shows the marginal propensity to consume out of disposable income is 0.5, the marginal rate is 0.5, and the marginal propensity to import is 0.3.
What does the slope of the aggregate demand schedule compare to the 45-degree line?

Explain your answer

The average marginal propensity to import means the cost of any increase in expenditure will be based on imports so that expenditure will exceed itself. The AD schedule is therefore steeper than the 45° line.

Your answer is correct. Review your answer to two decimal places.
By the midpoint is 0.11
It is possible to calculate the effect of an increase in the government spending on the aggregate demand schedule.
Instructions: On this graph, use the line tool (left), plot 2 points to draw the new aggregate demand curve.

Your Grade Score: 100%

connect

MATHS 3.1

MARKET EQUILIBRIUM WITH LINEAR DEMAND AND SUPPLY

We can describe the equilibrium in a given market in a simple mathematical way. First, we consider a **direct demand function** as a relationship between the quantity demanded and the price or service, keeping 'other things constant'. Suppose that the direct demand function is written as

$$Q^D = a - bP$$

where Q^D denotes the quantity demanded, P the price and a and b are two positive constants. Instead of numbers makes the analysis a bit more general. Equation (1) implies a negative relationship between the quantity demanded and the price of a given good or service.

Next, we introduce a linear **direct supply function**

$$Q^S = c + dP$$

where Q^S is the quantity supplied and c and d are two constants. We assume that the constant d is positive. This implies that there is a positive relationship between the quantity supplied and the price. The market equilibrium is where quantity demanded equals quantity supplied, meaning

ACTIVITY 1.1

SCARCITY IN ACTION: THE OLAGH EXPERIMENT

In 1945 R. A. Radford, a British economist, published a paper on the Oflag experiment during the Second World War. He explained how in a society like markets rose naturally to deal with the problem of scarcity. The story takes place in an Oflag, a camp for officers of the British Army who were imprisoned between 1943 and 1945. There was no private property; all the food parcels from the German Army and the Red Cross went to the prisoners. At the beginning the goods rations, such as chocolate, jam, etc., were distributed equally among the prisoners. There was no income inequality in that society. How did the society deal with scarcity? For example, at the beginning

CONCEPT 1.2

BEHAVIOURAL ECONOMICS

Rationality is a fundamental assumption in standard economic behaviour using that assumption. However in some cases it does not fit with observed human behaviour. Why do people appear to be risk-averse? Why do they like gaining it? Why do we sometimes overspend? Why do investors tend to express excessive optimism or pessimism? In those cases a different explanation that departs from the standard model is needed. This subject is a relevant branch of economics called behavioural economics. It studies the psychological and sociological influences on the individual's behaviour. Behavioural economics has become very popular over the last few years.

APPLICATION

Each chapter includes a number of boxed examples. These aim to show how a particular economic example can be applied in practice. They are explained in further detail on pages xviii–xxii. Connect also provides helpful tools to relate what is learnt in the chapter to real-world situations.

the book to real life, including additional case studies, a range of videos and animated graphs.

CASE 1.3

POOR MARX FOR CENTRAL PLANNERS

During the Cold War, economists used to argue about the merits of central planning. In the Soviet bloc, falling increasingly behind the living standards of the West, the planners were celebrating the arrival of their new owner, Roman Abramovich, in the Russian economy, initially as an oil trader and then as chairman of the state oil company. The Berlin Wall fell because the Soviet bloc had fallen behind the West. The difficulties that had emerged were:

- Information overload. Planners could not keep track of all the factors that affected production.

A **normal good** has a positive income elasticity of demand.

An **inferior good** has a negative income elasticity of demand.

A **luxury good** has an income elasticity above unity.

A **necessity** has an income elasticity below unity.

A 1 per cent income rise leads to a rise in quantity demanded but prices remain unaltered. The budget share of the good falls. Higher incomes and household budgets are spent at constant prices. Conversely, the budget share of necessities increases because the income elasticity of demand for luxuries is greater than 1 per cent. Rises in income reduce the budget share of necessities.

Inferior goods tend to be goods for which there exist more expensive substitutes.

In Chapter 3 we distinguished **normal goods**, **inferior goods**, for which demand falls as income rises, **luxury goods** and **necessities**. All inferior elasticities of demand are negative. However, the income elasticity of demand lies between zero and unity. These definitions tell us what happens to demand when income rises but prices remain unaltered. The budget share of the good falls. Higher incomes and household budgets are spent at constant prices. Conversely, the budget share of necessities increases because the income elasticity of demand for luxuries is greater than 1 per cent. Rises in income reduce the budget share of necessities.

The growth rate of US real GDP from 1983 to 2010 is shown in Figure 1.1. The graph shows significant fluctuations, with peaks around 1984, 1990, 1995, and 2000, and troughs around 1985, 1991, 2001, and 2009.

Figure 1.1 The growth rate of US real GDP: quarterly data
Source: Federal Reserve Bank of St. Louis.

AIDING UNDERSTANDING

Each chapter offers extensive pedagogy to aid your understanding of the topics. This includes Learning Outcomes, Key Terms, Summaries and extensive tables and figures. Further tools, such as summaries and

videos, are available in Connect.

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 define the relationship between utility and tastes for goods
- 2 describe the concept of diminishing marginal utility
- 3 describe the concept of diminishing marginal rate of substitution
- 4 understand how to represent tastes as indifference curves
- 5 understand how to derive a budget line
- 6 realize how indifference curves and budget lines expand or contract
- 7 describe how consumer income affects quantity demanded
- 8 understand how a price change affects quantity demanded
- 9 define income and substitution effects
- 10 recognize the market demand curve



Vignette

14.2 Growth and aggregate supply

Economic growth

Economic growth simply means an increase in real GDP between one year and the next.

Economic growth causes the long run AS curve to shift to the right such that higher AD can be met without increased inflation.

Sources of economic growth

There are 2 main sources of growth:

1. More factor inputs – through capital (physical and human) accumulation, an increase in the size of the labour force (through faster population growth, immigration and increased participation rates) and an increase in land and natural resources
2. Greater productivity – using resources more efficiently resulting in increased output per worker per man-hour (technical advance)



Boxed material

The text offers a wealth of boxed material to help explain how a particular economic principle can be applied in practice. There are four different types of box:

- **Case:** These draw on real-life companies and topical events to help illustrate economics in action.
- **Concept:** These contain content that is slightly more complex and provide some additional, more challenging, topics for you to explore.
- **Activity:** Similar to those in the last edition, these illustrate key economic concepts and then ask you to apply them to the real world. There are questions at the end of each activity and you can find the answers at the back of the book.
- **Maths:** These boxes highlight key mathematical formulae and present them in a digestible format. They can easily be skipped without interrupting the flow of the chapter, if necessary.

Chapter	Title	Type	Page
1	Opportunity costs and choices: an example	Concept	3
	Most output is service	Case	4
	The oil price shocks	Case	7
	Behavioural economics	Concept	10
	Scarcity in action: the rise of markets in a POW camp	Activity	14
	Poor Marx for central planners	Case	15
2	Hyperinflation	Concept	26
	Money illusion	Concept	28
	Landing the big job	Activity	33

	Get a Becker view: use an economist's spectacles	Case	34
3	Market equilibrium with linear demand and supply	Maths	44
	Horsemeat burger? No, thanks!	Case	46
	The demand for mobile apps	Case	48
	Movement along a curve vs shifts of the curve	Activity	49
	Graphical derivation of consumer and producer surplus	Concept	51
	Rent ceiling in Sweden	Case	53
	More on price controls	Case	54
	Anatomy of price and quantity changes	Case	56
4	Practising calculation of price elasticity of demand (PED) and the arc elasticity of demand	Maths	62
	The point elasticity of demand	Maths	64
	The price of the iPhone and the elasticity of demand	Case	68
	Easy profits	Case	71
	Car crazy	Case	75
	The point elasticity of supply	Maths	77
	The effects of a specific tax	Maths	79
5	Other contour maps	Concept	89
	The budget constraint and the budget line	Maths	91
	Do consumers really behave as utility-maximizing agents?	Concept	94
	Income and substitution effects in practice: the effects of petrol prices on grocery expenditure	Case	100
	Are cigarettes inferior goods?	Case	104
	Marginal activity and the water-diamond paradox	Concept	112
	Utility function, marginal utility and the marginal rate of substitution	Maths	114
6	The anatomy of a crisis: reading the balance sheet of Northern Rock	Case	119
	Economic vs accounting profits	Concept	121
	Hostile takeovers	Concept	121
	Marginal costs in practice: the case of local bus transportation in the UK	Case	126
	Total and marginal revenue with a linear demand	Maths	129
	The mathematics of profit maximization	Maths	132

7	The UK productivity puzzle	Case	143
	The short-run production function: the average and marginal product of labour	Maths	144
	Short-run cost functions	Maths	149
	Marginal conditions and sunk costs	Activity	151
	Scale economies and the Internet	Case	156
	The long-run production function and the returns to scale	Maths	157
	Globalization, technical change and income distribution	Case	159
8	Why do we need to study perfectly competitive markets?	Concept	172
	Profit maximization and monopoly power	Maths	182
	Monopoly power and competition policy	Case	186
	Regulation of natural monopolies	Concept	193
9	Market structure of the PC industry	Case	199
	It's not what it looks like	Concept	201
	Deriving the reaction functions in a Cournot duopoly	Maths	209
	Mergers and competition policy	Concept	211
	The Stackelberg model	Maths	212
	Barriers at the checkout	Case	216
10	The demand for inputs and profit maximization	Maths	225
	Individual labour supply and indifference curves	Maths	228
	The puzzle of low female participation in the labour force in Turkey	Case	232
	Does immigration hurt native workers?	Case	234
	How common is monopsony?	Concept	236
	Higher education pays off	Case	238
	Minimum wages hurt jobs, don't they?	Concept	239
11	The simple algebra of present values and discounting	Maths	253
	Factor markets: a summary	Concept	260
	The best address	Case	262
	How to measure income inequality: the Gini coefficient	Concept	264
12	Why play a losing game? The case of the National Lottery	Case	271
	Choice under uncertainty: expected utility of income and attitude to risk	Maths	273

	Why are CEOs paid so much?	Case	276
	Education and signalling	Concept	277
	Stock market volatility	Case	278
	Beta in action	Activity	284
	Excessive risk and the 2007-08 financial crisis: a behavioural view	Case	286
13	General vs partial equilibrium: an example from school policy	Concept	297
	Equity vs efficiency in trading: the Edgeworth box	Concept	299
	Externalities and the London 2012 Olympic Games	Case	305
	Internalizing a negative externality using property rights	Maths	308
	Stern view of discount rates	Activity	313
14	The paradox of open source software	Case	325
	Do you mind if I smoke? The smoking ban in the UK	Case	329
	Using a tax to internalize the negative externality	Maths	331
	Hunting the median voter	Activity	336
15	The circular flow revisited	Maths	351
	Problems in measuring GDP	Case	353
	Tax evasion, crime and under-reporting of GDP	Case	356
	Sustainability, pollution and negative GDP growth	Activity	359
16	A brief history of macroeconomics	Case	367
	Exogenous and endogenous variables	Concept	371
	The AD schedule: moving along it or shifting it?	Activity	373
	The Little Depression and the Great Depression	Case	374
	How did the crash affect investment?	Case	375
	Autonomous demand and equilibrium output	Maths	377
	Inflationary and deflationary gaps	Concept	379
	How stable is the saving rate?	Case	382
	Saving or savings?	Concept	384
17	Fiscal policy, austerity and debt: lessons from Japan	Case	391
	Cyclical output fluctuations and the government budget	Case	395
	The limits to fiscal policies	Concept	398
	Fiscal stability, responsibility and aggregate demand	Activity	401

	Equilibrium output revisited	Maths	408
	Globalization and international trade	Case	408
18	Barter economy vs monetary economy	Concept	415
	A beginner's guide to financial markets	Activity	417
	The money multiplier	Maths	420
	The collapse of bank lending	Concept	421
	The sub-prime crisis and its aftermath	Case	427
19	Preventing a future banking crisis?	Case	438
	Shifts in money demand	Maths	442
	Quantitative easing	Concept	443
	Quantitative easing revisited	Maths	446
	Transmission lag	Activity	450
	The credit channel of monetary policy	Concept	453
20	Fiscally challenged Eurozone countries	Case	459
	A modern interpretation of the LM schedule	Concept	462
	A horizontal LM schedule	Concept	464
	The monetary fiscal mix	Maths	466
	Monetary or fiscal policy?	Activity	467
21	Aggregate demand, the IS schedule and the ii schedule	Concept	476
	Supply shocks: oil prices and inflation	Case	479
	Output gaps, 1998–2013	Activity	485
	A Taylor rule for monetary policy	Concept	490
	The formula for the Taylor rule	Maths	490
22	The quantity theory of money: $MV = PY$	Concept	495
	Expectations formation and the accelerationist hypothesis	Concept	505
	Short-run and long-run Phillips curves	Maths	506
	Public enemy number two	Case	511
	Is Asia's inflation under control?	Activity	513
23	Measuring unemployment	Case	523
	The lump-of-labour fallacy	Concept	526
	Does the tax carrot work?	Activity	532
	The labour market incidence of policy changes	Maths	534

	How bad could unemployment become?	Case	535
	Hysteresis and high unemployment	Concept	538
24	Effective exchange rates	Concept	544
	International flows of financial capital	Case	552
	Interest parity conditions	Maths	554
	Changes in equilibrium real exchange rates	Activity	556
	Currency wars	Case	560
25	Reintroducing capital controls?	Case	566
	Aggregate demand under fixed exchange rates	Maths	573
	A 30-year look at sterling	Activity	578
26	The gold standard and capital flows	Concept	585
	A simple model of a balance of payments crisis	Maths	590
	World economy in a cul de sac?	Case	591
	Capital controls and the early ERM	Case	593
	Sovereignty and monetary union	Concept	596
	The economic consequences of Mr Draghi	Case	598
	Irish deflation and its consequences	Activity	601
27	The cyclical behaviour of wages	Activity	609
	The multiplier-accelerator model of cycles	Maths	613
	Eurozone business cycles	Case	615
	Dynamic stochastic general equilibrium models	Concept	618
28	Growth and competition	Case	636
	Neoclassical growth theory	Maths	640
	Aborted take-offs on the growth runway	Activity	642
	Does convergence occur in practice?	Case	644
29	Comparative advantage and the gains from trade	Concept	659
	The gains from trade	Maths	661
	Historical gainers and losers from trade	Case	663
	The EU Single Market	Activity	671

Online resources

Online Learning Centre

www.mcgraw-hill.co.uk/textbooks/begg



The Online Learning Centre accompanying this book provides a wealth of resources to assist you in your teaching. These are available to all adopters of the text.

- PowerPoint slides
- Lecture plans
- Teaching tips
- Tutorial exercises
- Discussion questions
- Case study teaching notes
- Solutions manual
- Suggested course structures

Testbank available in McGraw-Hill EZTest Online



A test bank of hundreds of questions is available to lecturers adopting this book for their module. A range of questions is provided for each chapter, including multiple-choice, true or false, and short-answer or essay questions. The questions are identified by type, difficulty and topic to help you to select questions that best suit your needs, and are accessible through an easy-to-use online testing tool, **McGraw-Hill EZ Test Online**.

McGraw-Hill EZ Test Online is accessible to busy academics virtually anywhere – in their office, at home or while travelling – and eliminates the need for software installation. Lecturers can choose from question banks associated with their adopted textbook or easily create their own questions. They also have access to hundreds of banks and thousands of questions created for other McGraw-Hill titles. Multiple versions of tests can be saved for delivery on paper or online through WebCT, Blackboard and other course management systems. When created and delivered through EZ Test Online, students' tests can be immediately marked, saving lecturers' time and providing prompt results to students. To register for this FREE resource, visit www.eztestonline.com

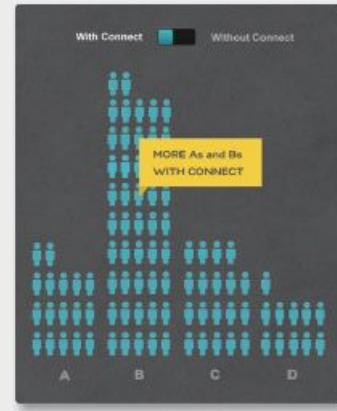


connect[®]

ECONOMICS

McGraw-Hill Connect Economics is a learning and teaching environment that improves student performance and outcomes whilst promoting engagement and comprehension of content.

You can utilize publisher-provided materials, or add your own content to design a complete course to help your students achieve higher outcomes.



PROVEN EFFECTIVE INSTRUCTORS

With McGraw-Hill Connect Plus Economics, instructors get:

- Simple **assignment management**, allowing you to spend more time teaching.
- **Auto-graded** assignments, quizzes and tests.
- **Detailed visual reporting** where students and section results can be viewed and analysed.
- Sophisticated **online testing** capability.

- A **filtering and reporting** function that allows you to easily assign and report on materials that are correlated to learning outcomes, topics, level of difficulty, and more. Reports can be accessed for individual students or the whole class, as well as offering the ability to drill into individual assignments, questions or categories.
- **Instructor materials** to help supplement your course.

Get Connected. Get Results.

STUDENTS

With McGraw-Hill Connect Plus Economics, students get:

Assigned content

- Easy **online access** to homework, tests and quizzes.
- **Immediate feedback** and 24-hour tech support.

Self-Quiz and Study

- **Practice tests** help you to easily identify your strengths and weakness and create a clear revision plan.
- A fully searchable **e-book** allows you to brush up on your reading.
- **Study tools** – including videos, animated graphs, vignettes and summaries give you extra practice on individual topics.

The screenshot displays the Connect platform interface. At the top, there is a navigation bar with tabs for Home, Library, and Reports. Below the navigation bar, a large section is titled "Self Study Course" with a "Self-Study" button. This section contains a graphing question labeled "13. Value Consumer's Income is \$30. Price units of X per unit and Price units of Y per unit." The question asks students to plot a budget line and find the optimal consumption point. A graph is provided with axes for Price (Y-axis) and Price (X-axis), showing two intersecting lines. Below the graph, instructions and a note about saving work are given. To the right of the graph, there is a "library" section showing a grid of chapter items from Chapter 1 to Chapter 6, each with a title, a green progress bar, and a "View Details" link.

Self Study

If your instructor is **not** prescribing Connect as part of your course, you can still access a full range of student support resources on our Self Study platform at <http://connect.mcgraw-hill.com/selfstudy>.

ACCESS OPTIONS



Is an online assignment and assessment solution that offers a number of powerful tools and features that make managing assignments easier, so faculty can spend more time teaching. With Connect Economics, students can engage with their coursework anytime and anywhere, making the learning process more accessible and efficient.

Graphing tools

Enable students to develop their graphical ability, calculating and plotting data as part of basic and complex questions. Auto-graded graphing questions are available throughout the assignable and student selfstudy content and can provide immediate

feedback. Students can also access animated graphs as part of their Study Tools, giving them the chance to visualize and understand key concepts step by step.

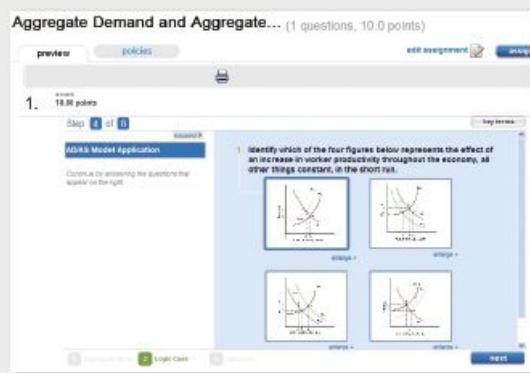
Calculation questions

Test students' mathematical understanding with auto-graded calculation questions.



Logic cases

Assign multi-part problems that cover key topics in economics and then branch to different follow-up questions, activities and analysis.

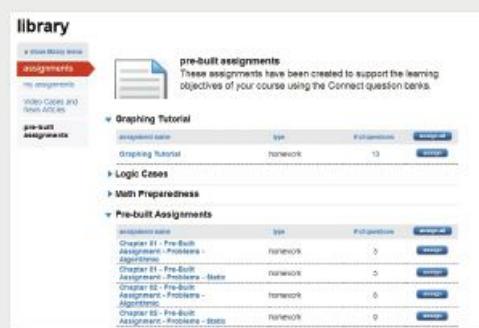


Algorithmic problem sets

Provide repeated opportunities for students to practise and master concepts with multiple versions of each problem. Or use the algorithmic problems in class testing to provide each student with a different version than that seen by their peers.

Short-answer questions

Ensure students develop strong writing skills with shortanswer and essay questions. Each question provides a guide answer and allows you to review and mark student responses. These questions are clearly marked as being manually graded, so you can include or skip these as you see fit.



Pre-built assignments

Assign all of the end of chapter or test bank material as a ready-made assignment with the simple click of a button.



Seamlessly integrates all of the Connect Economics features with:

- An integrated e-book, allowing for anytime, anywhere access to the textbook.
- Dynamic links between assigned and self-study problems and the location in the e-book where that problem is covered.
- A powerful search function to pinpoint and connect key concepts.

e-book

Chapter 14. Supply-side policies and economic growth

Business problem: assessing economic growth

The economic growth rates for almost European Union (EU) members and the USA from 1990 to 2011. All economies have experienced periods of slow growth. All economies slowed between 2007 and 2009, reflecting the impact of the financial events of 2007 on the global economy. After 2009, most economies grew well until 2011, except for Ireland which had a sharp decline. In many economies there was a brief slowdown after 2011, followed by recovery. Over the period and across economies, average growth seems to have been around 2 per cent per annum.

Figure 14.1 GDP growth rates for various economies

Practice quiz

Missed questions

Most challenging learning objectives

Least on Demand

Current learning status

Self-Assessment

Time of Knowledge

According to the law of demand, as market price increases, quantity demanded _____

Do you know this answer? [See below]

Yes Probably Maybe No—Just guessing

LearnSmart™

McGraw-Hill LearnSmart is an adaptive learning program that identifies what an individual student knows and doesn't know. LearnSmart's adaptive learning path helps students learn faster, study more efficiently, and retain more knowledge. Reports available for both students and instructors indicate where students need to study more and assess their success rate in retaining knowledge.



create

Make our content your solution

At McGraw-Hill Education our aim is to help lecturers to find the most suitable content for their needs delivered to their students in the most appropriate way. Our **custom publishing solutions** offer the ideal combination of content delivered in the way which best suits lecturer and students.

Our custom publishing programme offers lecturers the opportunity to select just the chapters or sections of material they wish to deliver to their students from a database called CREATE™ at

www.mcgrawhillcreate.co.uk

CREATE™ contains over two million pages of content from:

- textbooks
- professional books
- case books – Harvard Articles, Insead, Ivey, Darden, Thunderbird and BusinessWeek
- Taking Sides – debate materials

Across the following imprints:

- McGraw-Hill Education
- Open University Press
- Harvard Business Publishing
- US and European material

There is also the option to include additional material authored by lecturers in the custom product – this does not necessarily have to be in English.

We take care of everything from start to finish in the process of developing and delivering a custom product to ensure that lecturers and

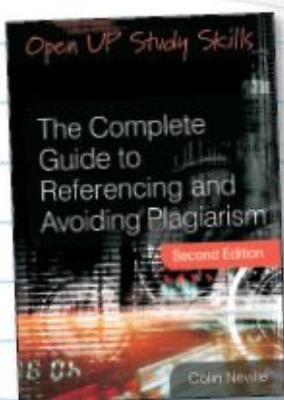
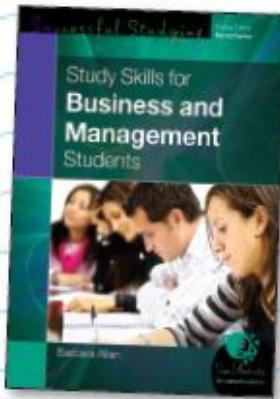
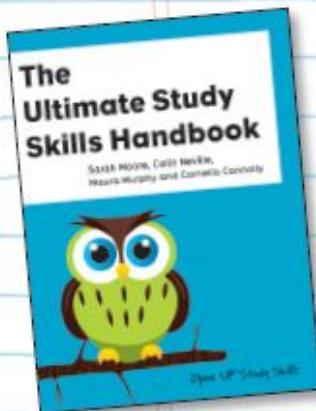
students receive exactly the material needed in the most suitable way.

With a **Custom Publishing Solution**, students enjoy the best selection of material deemed to be the most suitable for learning everything they need for their courses – something of real value to support their learning. Teachers are able to use exactly the material they want, in the way they want, to support their teaching on the course.

Please contact **your local McGraw-Hill representative** with any questions or alternatively contact Warren Eels e:
[http://www.warren_eels@mcgraw-hill.com](mailto:www.warren_eels@mcgraw-hill.com).

**Improve
your
grades!**

**20% off any
Study Skills
book!**



Our Study Skills books are packed with practical advice and tips that are easy to put into practice and will really improve the way you study.

Our books will help you:

- ✓ Improve your grades
- ✓ Write confidently
- ✓ Avoid plagiarism
- ✓ Undertake research projects
- ✓ Save time
- ✓ Sail through exams
- ✓ Develop new skills
- ✓ Find the perfect job

Special offer!

As a valued customer, buy online and receive 20% off any of our Study Skills books by entering the promo code **BRILLIANT!**

www.openup.co.uk/studyskills



Acknowledgements

Firstly, we would like to thank Kiera Jamison, Natalie Jacobs and all the other people at McGraw-Hill. A special thanks goes to our students at Imperial College London and the University of Essex for the useful feedback they have given us over the years.

We would like to thank the following reviewers who provided helpful suggestions and comments on the book as it progressed through its revision:

Jorn Rattso, Norwegian University of Science and Technology

Alan Reeves, University of the West of Scotland

Nicole Tabasso, University of Surrey

Paul Middleditch, University of Manchester

John Hudson, University of Bath

John Sessions, University of Bath

Martin Jensen, Birmingham University

Javier Valbuena, University of Kent

Dimitrios Syrrakos, Manchester Metropolitan University

Robert Butler, University College Cork

Gary Cook, University of Liverpool

Surja Datta, Oxford Brookes University

Jim Johnston, University of the West of Scotland

Andrea Mannberg, Umeå University

Aileen Murphy, University College Cork

Jennifer Roberts, University of Sheffield

Panos Sousounis, University of Keele

Finally, special thanks to the following academics who helped to produce some of the many online resources available with this text:

Jo Evans, University of Surrey
James Johnston, University of the West of Scotland
Sanna Nieminen, Jyväskylä University of Applied Sciences
Alan Reeves, University of the West of Scotland
Abhijit Sharma, Bradford University
Dimitrios Syrrakos, Manchester Metropolitan University
David Begg and Gianluigi Vernasca

October 2013

Every effort has been made to trace and acknowledge ownership of copyright and to clear permission for material reproduced in this book. The publishers will be pleased to make suitable arrangements to clear permission with any copyright holders whom it has not been possible to contact.

About the authors



Professor David Begg is Professor of Economics at Imperial College Business School.

Born in Glasgow, David went to Cambridge in the hope of playing cricket for England but became fascinated with economics. After also studying at Oxford, he won a Kennedy Scholarship to the Massachusetts Institute of Technology, where Stanley Fischer and Rudiger Dornbusch were his PhD supervisors.

An expert on monetary and exchange rate policy, David has advised the Bank of England, HM Treasury, the IMF and the European Commission. He is a fellow of the Royal Society of Edinburgh, a fellow of the City and Guilds of London Institute and a non-executive director of Imperial Innovations, which invests in technology start-ups from Imperial College, University of Cambridge, University of Oxford and UCL.

David has always been committed to showing how useful economics is in making sense of the world around us. His other books include *Foundations of Economics* (now in its fifth edition) and *Economics for Business* (co-authored with Damian Ward, now in its fourth edition).



Dr. Gianluigi Vernasca is Senior Lecturer in Economics at the University of Essex. Since October 2009 he has also been Director of Undergraduate Studies in the Department of Economics.

Gianluigi received his PhD in 2006 from the University of Warwick. His research is mainly in the field of industrial economics. He has taught economics at both undergraduate and postgraduate levels in various institutions. Gianluigi has also worked as an economic consultant on competition and antitrust issues.

For those of you wondering about David and Gianluigi's co-authors: after leaving MIT **Stanley Fischer** became Chief Economist of the World Bank, Deputy Head of the IMF, Vice Chairman of Citigroup and Governor of the Bank of Israel. Until his untimely death, **Rudiger Dornbusch** remained a professor at MIT; his analysis and recommendations were sought by countless governments and corporations. Stan and Rudi taught a generation of students – including Ben Bernanke, current Chairman of the US Federal Reserve, and Olivier Blanchard, current Chief Economist of the IMF.

PART ONE

Introduction

Economics is all around you. It is about how society deals with the problem of scarcity. We cannot have everything we want, whether this refers to continuous holidays or perfectly clean air. We have to make choices. Economics is the study of how society makes these choices. Economics is not just about incomes, prices and money. Sometimes it makes sense to use markets, sometimes we need other solutions. Economic analysis helps us decide when to leave things to the market and when to override the market.

Chapter 1 introduces the central issues of scarcity and choice, and the extent of government involvement in these decisions. Chapter 2 outlines economic reasoning, discussing how our understanding is advanced by the interaction of theories and evidence. Chapter 3 illustrates markets in action.

Contents

- 1** Economics and the economy
- 2** Tools of economic analysis
- 3** Demand, supply and the market

CHAPTER 1

Economics and the economy

Learning outcomes

By the end of this chapter, you should be able to:

- 1 understand that economics is the study of how society resolves the problem of scarcity
- 2 describe ways in which society decides what, how and for whom to produce
- 3 understand the concept of opportunity cost
- 4 differentiate between positive and normative economics
- 5 define microeconomics and macroeconomics

Economics is the study of how societies make choices under conditions of scarcity.

Scarcity means that we have limited resources to produce all the goods and services that we would like to have. Goods are physical commodities, such as steel or strawberries. Services are activities such as massages or live concerts, consumed or enjoyed only at the instant they are produced.

The key economic problem for a society is how to reconcile the conflict between people's virtually limitless desires for goods and services, and the **scarcity of resources** (labour, machinery, raw materials, and so on) with which these goods and services can be produced.

By emphasizing the role of society, our definition places economics within the social sciences that study and explain human behaviour.

Because of scarcity we cannot get everything we want. Choosing to spend more resources to build a new military ship means having fewer resources to build new schools or hospitals instead. Therefore a society faces *trade-offs* every time it makes choices. The three basic trade-offs that every society faces are: *what* goods and services to produce, *how* to produce them and *for whom* to produce them. In answering what, how and for whom to produce, economics explains how scarce resources are allocated among competing claims on their use.

A **resource** is **scarce** if the demand of that resource at a zero price would exceed the available supply.

Economics is the study of how society makes choices under conditions of scarcity.

In many societies the answer to those questions is the result of the independent decisions of many individuals and firms. Individuals decide how much to work, what to buy and how much to buy of various goods and services. Firms decide what to produce, how much to produce of various goods and services and which resources to use in production. Normally those decisions take place in markets. In economics we study how individuals and firms make decisions and how those decisions are combined in markets to determine how resources are allocated.

Not all resources are allocated through markets, though. For example, we may all want an unpolluted environment but there is no market for such a good. In some other cases the allocation created by markets may not be what a society would like. When this happens, there is scope for governments to intervene in the economy.

Although economics is about human behaviour, we describe it as a science. This reflects the method of analysis, not the subject matter, of economics. Economists develop theories of human behaviour and test them against the facts. Chapter 2 discusses the tools that economists use and explains the sense in which this approach is scientific. This does not mean that economics ignores people as individuals. Moreover, good economics retains an element of art. Only by having a feel for how people actually behave can economists focus their analysis on the right issues.

1.1

How economists think about choices

How do societies choose to allocate scarce resources among different alternatives? To answer that question we need to understand how each member of a society, such as individuals and firms, makes economic choices.

A fundamental assumption in economic analysis is that individuals behave *rationally* when making choices. A rational individual is someone who takes full account of all information available to make the best choice for his own interest. In taking decisions a rational individual uses all his available information to compare the benefits and the costs associated with those decisions. Rationality in making choices implies that a given activity should be undertaken only if its benefits are larger than its costs.

This idea of a rational choice provides an answer to the questions what to produce, how to produce and for whom to produce. We should produce something for which the benefits of producing it are larger than the costs of producing it. We should produce goods in the least costly way and for the individuals who value those goods most.

The benefits of a given activity are what we gain from doing it. In many cases those benefits can be measured in monetary terms. This is the amount of money we get from a given activity. The benefit you receive from a university degree is the boost in your earnings once you graduate. However, in many other cases providing a monetary measure for benefits is less straightforward. For example, suppose we need to choose whether or not to watch the last *Spiderman* film. What is the monetary benefit of watching that film?

Even when benefits cannot be readily available in monetary terms we can still try to provide a monetary value for them. Economists define the benefits of a given action as the willingness to pay for that action. The willingness to pay for something is the maximum amount of money we are willing to pay for it. This is a hypothetical monetary value; we normally do not pay that money. In the case of the last *Spiderman* film, you can ask yourself the hypothetical question, ‘What is the maximum amount of money I am willing to pay to watch that film?’ If you say, for example, £25, then this value should represent your benefit of watching that film. Notice that different individuals can have different willingness

to pay for the same thing. A real fanatic of *Spiderman* films will be willing to pay a higher amount of money to watch those films than someone who prefers to watch romantic comedies.

What about the costs of a given activity? The cost of an activity normally includes the *explicit cost*, which is the amount of money we need to pay for doing it, and the **opportunity cost**, which is the value of the best alternative we must sacrifice for doing that activity.¹ Therefore the cost of any activity generally has two components. A rational individual must consider both explicit and opportunity costs in making decisions.

The **opportunity cost** of an activity is the value of the best alternative you must sacrifice.

As for benefits, different individuals may have different opportunity costs for the same activity simply because they value different best alternatives.

The concept of opportunity cost is one of the most important concepts in economic analysis. It provides the direct link between scarcity and choices. Faced with scarcity, we must choose among alternative activities. If we decide to do activity A, then we are spending resources on that activity. We could have spent those resources on doing activity B. Therefore by choosing one activity we sacrifice the other. In making choices we need to take into account the opportunity that we are sacrificing.

CONCEPT 1.1

OPPORTUNITY COSTS AND CHOICES: AN EXAMPLE

To understand how opportunity costs play a crucial role in determining how individuals make choices, we can consider a simple example.

Suppose you want to spend a day at the beach with your friends. The cost of spending the day at the beach with your friends is £30 (this includes the cost of transportation, the cost of food and drink and so on). Suppose that you really like spending a day with your friends and that you will be willing to pay £60 to do so. This is the

maximum amount you are willing to pay for that activity and it represents its benefit in monetary terms.

Given these two facts, should you go to the beach with your friends? At first sight the answer appears to be yes. The benefit is larger than the cost. However, the cost of going to the beach is not only £30. By deciding to go to the beach you are losing the opportunity of doing something else. Suppose that the best alternative of going to the beach is to work at the campus shop at your university. A day of work at the campus shop gives you £70. This represents the benefit, in monetary terms, of working at the campus shop. Suppose you like working at the campus shop but only if they pay you at least £30. This is the minimum payment that will make you work at the shop. If they pay you less than £30, then you will be unwilling to work there. What is the value of working at the campus shop for you? The answer is simply £70 minus £30; that is, £40.

If faced with the choice between ‘going to the beach with your friends’ and ‘working at the campus shop’, you need to take into account that by choosing one alternative you are giving up the other. Therefore the cost of spending the day at the beach is not only the explicit cost of £30; it also reflects the implicit cost given by the best alternative you are sacrificing. In this case, the value of the best alternative is £40. This represents the opportunity cost of ‘going to the beach with your friends’. If we take into account this opportunity cost, then the cost of ‘going to the beach with your friends’ is £30, which is the explicit cost, plus £40, which is the opportunity cost, meaning it is £70. Given that the benefit of ‘going to the beach with your friends’ is £60 and it is lower than the cost (explicit cost plus the opportunity cost), as a rational individual you should choose not to do it; instead you should work at the campus shop.

While this example appears very simple, the basic principle behind it is very general. Every individual, firm, government and, therefore, every society, when making choices must compare the benefits and costs of those choices. Failing to take into account opportunity costs in our decision process will result in wrong choices.

You may now think that the way in which economists describe how individuals make choices is not very realistic. Probably you do not make explicit calculations about opportunity costs when you make choices. Generally you do not ask yourself how much you are willing to pay to do something. Nevertheless it turns out that, even if individuals do not make explicit calculations about benefits and costs every time they make choices, they behave in many cases as if they do. Therefore our way of thinking about individuals—choices can help us explain many facts about their economic behaviour.

Choices and incentives

We have seen that rational individuals compare benefits and costs (including opportunity costs) in making decisions. Another important aspect of economic analysis is that in making choices rational individuals respond to *incentives*. Incentives are rewards or penalties that affect individuals—choices. Incentives can be monetary or not and can affect the benefits or the costs of our decisions. For example, smokers may have an incentive to smoke less if the government imposes high taxes on the price of cigarettes. In this case the high tax on cigarettes affects the cost of smoking and thus the choice of smoking. Recently the UK government has increased the tuition fees for undergraduate degrees from £3000 to £9000. This increase in tuition fees can affect the incentives of individuals to enrol on undergraduate degrees in the UK.

Incentives are important to help us understand how markets work. The prices of goods and services create incentives that affect the decisions of individuals and firms. As the price of beef increases, buyers have an incentive to buy less beef and maybe to switch to a cheaper kind of meat, like chicken. On the other hand, beef producers have an incentive to produce more since the price of their product has increased.

CASE 1.1

MOST OUTPUT IS SERVICE

Societies must decide what to produce. This represents one of the basic three trade-offs societies face. The table below shows what is produced by different economies in 2010.

In advanced countries, agriculture represents about 1 per cent of national output and industry less than 25 per cent. The rest is services, which include banking, transport, entertainment, communications, tourism and public services (defence, police, education, health). In countries such as China and India, agriculture remains a higher share of GDP and services are not yet fully developed. Everywhere, services are the fastest growing part of output and of exports. Success in exporting banking, fashion and entertainment helps make the UK the second-largest exporter of services in the world.

In developed countries, services have for a long time been the largest component of national output. Until recently most international trade was trade in goods. The Internet has changed all that. Accounting services can be outsourced to India and the advice of Indian accountants is as rapidly received by email in the UK as face to face in India.

% of national output	UK	US	France	China	India
Agriculture	1	1	2	10	19
Industry	22	20	19	46	26
Services	77	79	79	44	55

Source: World Bank, *World Development Indicators*. © 2012 The World Bank Group. All rights reserved.

1.2 Economic issues

Trying to understand what economics is about by studying definitions is like trying to learn to swim by reading an instruction manual. Formal analysis makes sense only once you have some practical experience. In this section we discuss two examples of how society allocates scarce resources between competing uses. In each case we see the importance of the questions what, how and for whom to produce.

The 2007 financial crisis

The crisis that started in the US in 2007 is now considered to be the worst economic crisis since the 1929 Great Depression. Because of the

crisis many economies entered a recession. A *recession* is a period of time (at least two consecutive quarters) in which the amount of goods and services produced by an economy (called the gross domestic product, or GDP) declines. Recessions are not uncommon; indeed, economies over time experience cyclical periods of recession followed by periods in which economic activity rises.

Figure 1.1 shows the GDP growth rate for the US economy from 1980 to 2012. It shows that in the last 22 years there have been several periods in which the US economy faced a negative growth rate in terms of output; however, none to the extent of the recession that began in 2007.

The situation started with a financial crisis in the US market for sub-prime mortgages. Sub-prime mortgages are those given to people who want to buy a house but have a poor credit record and a relatively high probability of eventually not being able to repay their loan.

Why lend money to people who are such a high credit risk? The reason was the constantly increasing price of houses before 2007, as shown in Figure 1.2. With the price of houses increasing, the risk of losing on a sub-prime mortgage was limited. The increase in house prices created an incentive for banks to lend to sub-prime customers. In the case of default on a mortgage by a sub-prime borrower, the lender could repossess their house and sell it at a high price. To make those subprime mortgages more profitable, mortgage companies started to bundle them into bonds and sell them to other financial companies. By doing so, the mortgage companies were in effect borrowing to lend mortgages.



Figure 1.1 The growth rate of US real GDP: quarterly data

Source: Federal Reserve Bank of St. Louis.

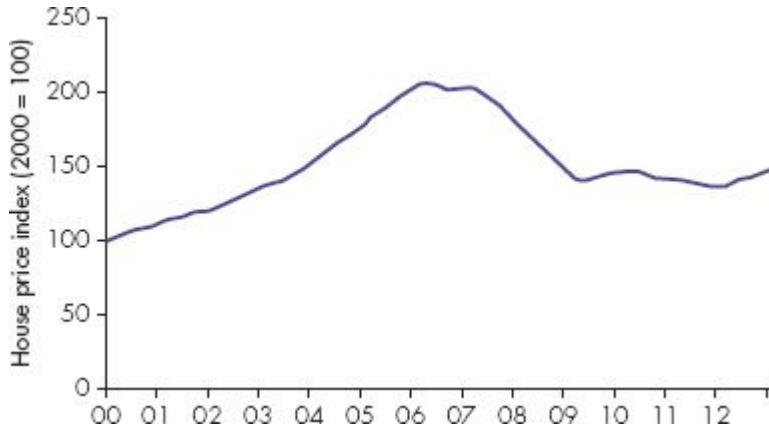


Figure 1.2 The US house price index: monthly data

At the same time, the interest rate on mortgages was relatively low in the US, creating an incentive for people to get a mortgage and buy a house.

The sub-prime mortgage market in the US is relatively small, comprising around 10 per cent of the entire market. How can such a small market create such a huge effect? After 2002, bond issuance increased substantially, as shown in Figure 1.3.

The problem started when more and more sub-prime borrowers started to default on their mortgages and repossession increased. The wave of repossession had a dramatic effect on house prices. More houses were available in the market, which caused the house price boom of the previous few years to reverse.

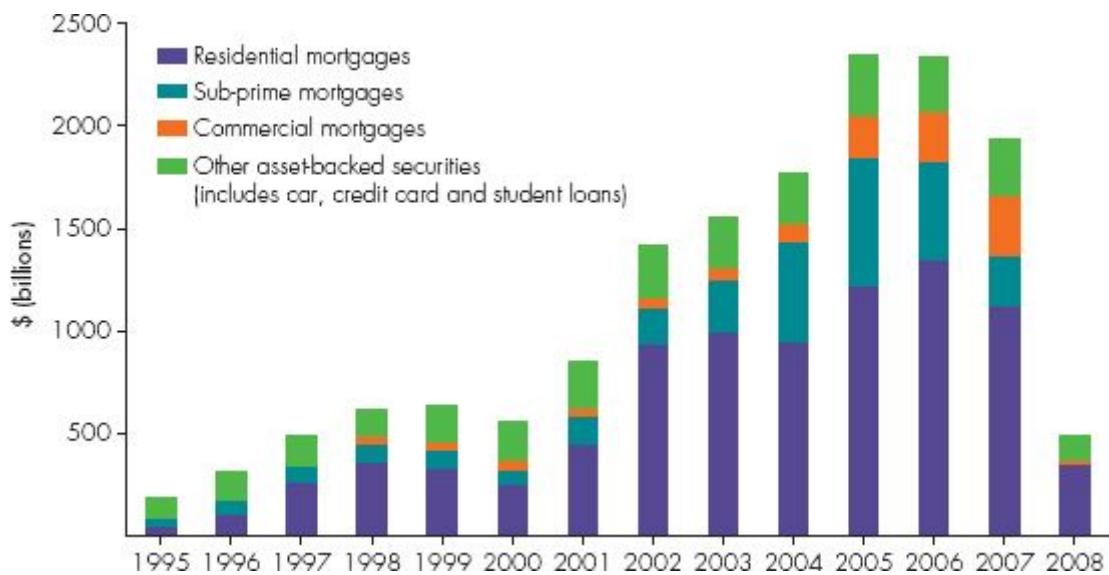


Figure 1.3 Global issuance of bonds backed by mortgages

Source: Bank of England. © 2010 The World Bank Group. All rights reserved.

With decreasing house prices, the sub-prime mortgages started losing value. Those financial investors who had bonds with sub-prime mortgages started experiencing big financial losses. At the end of 2007 many banks announced \$60 billion worth of losses, as many of the bonds backed by sub-prime mortgages had fallen in value.

The financial crisis started in the US. However, given the connectedness of financial institutions across countries, it rapidly spread globally. Governments in different countries started bailout programmes to avoid many big banks and financial companies collapsing and disappearing. Those companies were considered too big to fail.

The losses made by the financial and banking sectors resulted in what is called a *credit crunch*. A credit crunch is a reduction in the availability of credit from the banking sector. Banks become more reluctant to lend to borrowers and, more importantly, to each other. Credit becomes a scarce resource. The contraction of credit affected economic activity in all sectors of various economies and this had the effect of worsening the crisis.

Many firms rely on borrowing for running their usual business. The scarcity of credit available caused those firms to reduce their activity. Many closed down; others reduced their activity by cutting jobs.

Unemployment started to rise. Consumers—confidence started to decrease. People started to save more and to spend less.

For many economies the recession caused by the financial crisis started at the end of 2007/beginning of 2008. The first signs of a slight recovery appeared at the end of 2009. However, five years after the credit crunch started, most developed economies have still not fully recovered.

The financial crisis has affected the way societies answer the questions of what, how and for whom to produce. Production decreased because of the recession. Some sectors were more affected than others. The building sector suffered big losses as a result of falling house prices and the number of new houses built fell substantially during the crisis. Expensive and non-essential goods also suffered. Demand for cars fell and in many cases governments intervened by providing incentives to buy new cars to sustain demand. However, investments in research and development (R&D) and software increased.

Looking at how things are produced, the crisis created a reallocation of input resources used in production. This reallocation has affected labour input most dramatically. Reduction in economic activity implies a reduction in employment. The job market shrinks and so more people become unemployed. Moreover, finding jobs becomes even more difficult, causing unemployment to last longer.

Finally, the ‘for whom’ question. As a result of the crisis the banking and financial sector suffered big losses. The same happened to many investors because of the fall in stock values in financial markets. The crisis has hurt middle- and low-income households more than the very rich ones. The fall in the price of houses made firsttime buyers better off. Discount supermarkets’ sales figures also rose.

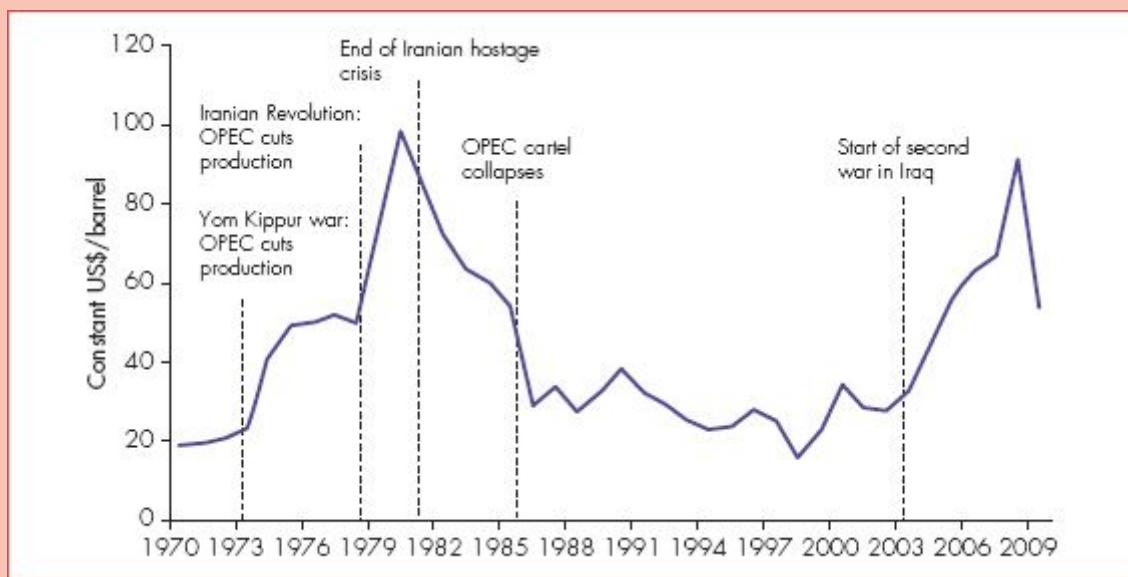
CASE 1.2

THE OIL PRICE SHOCKS

Oil provides fuel for heating, transport and machinery and is an input for petrochemicals and household products ranging from plastic plates to polyester clothes. What happens if continuing uncertainty in the Middle East or the ravages of climate change lead to very high oil prices? A little history lesson is useful in thinking about the likely results.

Up to 1973 the use of oil increased steadily. It was cheap and abundant. In 1973 OPEC – the Organization of Petroleum Exporting Countries (www.opec.org) – organized a production cutback by its members, making oil so scarce that its price tripled. Users could not quickly do without oil. Making oil scarce was very profitable for OPEC members.

The figure below shows the real (inflation-adjusted) price of oil, measured in US dollars, from 1970 to 2009. The price tripled between 1973 and 1977, doubled between 1979 and 1980, but then fell steadily until the mid-1990s. Markets found ways to overcome the oil shortage that OPEC had created. High oil prices did not last indefinitely. Given time, the higher price induced consumers to use less oil and non-OPEC producers to sell more. These responses, guided by prices, are part of the way many societies determine what, how and for whom to produce.



Consider first *how* things are produced. When the price of oil jumps, firms cut their use of oil-based products. Chemical firms develop artificial substitutes for petroleum inputs, airlines order more fuel-efficient aircraft, and electricity is produced from more wind farms. Higher oil prices make the economy produce in a way that uses less oil.

How about *what* is being produced? Households install better insulation to economize on expensive central heating and they buy smaller cars. Commuters form car-pools or move to the city centre. High prices choke off the demand for oil-related commodities but

also encourage consumers to purchase substitute commodities. Higher demand for these commodities bids up their price and encourages their production. Designers produce smaller cars, architects use solar energy and research laboratories develop alternatives to petroleum in chemical production.



© Instinia | Dreamstime.com

The *for whom* question in this example has a clear answer. The revenues of oil producers increased sharply. Much of their increased revenue was spent on goods produced in the industrialized Western nations. By contrast, oil-importing nations had to give up more of their own production in exchange for oil imports. In terms of goods as a whole, the rise in oil prices raised the buying power of OPEC and reduced the buying power of oil-importing countries such as Germany and Japan. The world economy was producing more for OPEC and less for Germany and Japan.

After 1982, OPEC's power diminished as other oil supplies came on stream and users developed adequate substitutes. However, OPEC got its act together again in 1999, cut supply, forced up oil prices and prompted another fuel crisis in 2000.

Since 1999 the sharp rises in oil prices must also be attributed not merely to a restriction of oil supply but also to a surge in energy demand by emerging economies, particularly China and India.

The oil price shocks show how society allocates scarce resources between competing uses. The higher oil price reflected its greater scarcity when OPEC reduced production levels.

Source: adapted from *The Economist*, 21 July 2007.

Income distribution

You and your family have an annual income that lets you enjoy various goods and services and live in a particular place. Your standard of living includes the necessities of life – food, shelter, health, education – and something beyond, such as recreation. Your income is lower than some people's but higher than that of others.

Nations also have different levels of income. A nation's income, or national income, is the sum of the incomes of all its citizens. World income is the sum of all countries' incomes, hence also the sum of the incomes earned by all the people in the world.

Income distribution is closely linked to the what, how and for whom trade-offs. Table 1.1 shows the percentage of the world population that lives in different groups of countries. Twenty-one per cent of the world's population lives in poor countries, such as Bangladesh and Indonesia. Seventy per cent live in middle-income countries, a group including Thailand, Brazil, Mexico and China. The rich countries, including the US, Western Europe, Canada and Japan, account for 9 per cent of the world population.

The **income distribution** (in a country or in the world) tells us how total income is divided between different groups or individuals.

Table 1.1 World population and income

	Country group		
	Poor	Middle	Rich
Income per head (£)	300	3900	17700
% of world population	21	70	9
% of world income	3	19	78

Source: World Bank, *World Development Indicator*. © 2010 The World Bank Group. All rights reserved.

Income per person indicates the average standard of living. Table 1.1 shows that in poor countries the average income per person is only £300 a year. In the rich industrial countries annual income is £17 700 per person, nearly 60 times larger. These are big differences.

Table 1.1 also shows that poor countries account for one-fifth of the world's population but only 3 per cent of world income. Rich countries have 9 per cent of the world's population but 78 per cent of world income. For whom does the world economy produce? Mainly for the 9 per cent of its population living in the rich industrial countries. This also helps answer what is produced. World output is directed mainly to the goods and services consumed in the rich countries. The degree to which income is unequally distributed within a country affects also which goods are produced. In countries where income is very unequally distributed, the rich can employ low-income people as maids, cooks and chauffeurs. In countries where equality is much greater, few people can afford to hire servants.

Why is inequality so great? This reflects mainly how goods are produced. Poor countries have little machinery and few people with professional and technical training. One American worker uses power driven earth-moving equipment to complete a task undertaken in Africa by many more workers equipped only with shovels. Workers in poor countries are less productive because they work under adverse conditions.

Income is unequally distributed within each country as well as between countries. In Brazil, the richest 10 per cent of families get 48 per cent of national income, but in Denmark, only 20 per cent.

These differences partly reflect things like state education, which increases access to education and training. However, in looking at income distribution within a country, we must include two extra things that are often less important when discussing differences in income per person between countries.

First, individual incomes come not just from working but also from owning assets (land, buildings, corporate equity) that earn rent, interest and dividends. In Brazil, ownership of land and factories is concentrated in the hands of a small group; in Denmark, it is not.

Second, societies may decide whether to change their distribution of income. A pure socialist economy aims to achieve considerable equality of income and wealth. In contrast, in an economy based on private ownership, wealth and power become concentrated in the hands of a few people. Between these extremes, the government may levy taxes to alter the income distribution that would otherwise emerge in a private ownership economy.

Income inequality also affects individuals— incentives. Low-income people have little incentive to save and a higher incentive to borrow to increase their consumption. If governments decide to increase taxes to smooth income inequality, this affects the incentive of the high-income individuals. In 2012 France proposed introducing a 75 per cent income tax on the very rich. Introducing such a tax would create an incentive for the very wealthy French to move their business to countries where taxes are lower.

CONCEPT 1.2

BEHAVIOURAL ECONOMICS

Rationality is a fundamental assumption in standard economics. We can explain a great deal of individuals' behaviour using that assumption. However, in some cases rationality seems to fail to accurately explain observed human behaviour. Why do people appear to dislike losing something roughly twice as much as they like gaining it? Why do we sometimes overspend on things we do not really need? Why in certain periods do investors tend to express excessive optimism about the value of stock prices?

In those cases a different explanation that departs from rationality should be considered. This is the main subject of a relevant branch of economics called behavioural economics. Behavioural economics is the study of psychological and sociological influences on the individual's behaviour.

Behavioural economics has become very popular over the years, especially after the recent financial crisis when it appeared that some investors may have underestimated the risk of their investments.

The assumption of rationality implies that people take full account of all information available to make their decisions. What are possible departures from this assumption? Here are some possibilities.

Suppose there is a fixed cost of either acquiring information or of taking the time to make a decision. This leads to *bounded rationality*. It is no longer optimal to examine every possible decision in great detail – you would incur too many fixed costs – so instead you incur costs once, have a good think, and then come up with a simple decision rule that you implement automatically until it no longer fits the facts, at which point you incur some more thinking costs and try to improve your rule. Simple rules may explain why people extrapolate the recent past rather than conduct extensive research all the time.

In 1979 Kahneman and Tversky, two psychologists, proposed *prospect theory* as a means to explain why individuals in some circumstances appeared to be loss averse: they are more sensitive to losses compared to gains of similar magnitude. According to this theory, individuals have different preferences for losses compared to gains mainly for psychological reasons. In 2002 Daniel Kahneman was awarded the Nobel Prize in Economics.

Richard Thaler, an economist at the University of Chicago, proposed *nudge theory* to show how psychological positive reinforcement can alter individuals' decisions. Here is an example: in the US two letters were sent to non-payers of vehicle taxes. First, a generic letter was sent to non-payers reminding them that they would lose their car if they didn't pay the tax. Second, a more personalized letter was sent with an attached picture of the car in question. The rewritten letter tripled the number of people paying the tax.

One possible criticism of behavioural economics is that, while there is only one way to be rational, there are millions of ways in which to be irrational. Anyone can explain a particular event by invoking a particular kind of irrationality – it then takes a lot of data to establish whether there is anything systematic in this irrationality or whether it was just a coincidence invoked by someone trying to be wise after the event.

1.3 Scarcity and the competing use of resources

In this section we consider a simple economic model to show how a society chooses to allocate scarce resources between competing uses.

Consider an economy with workers who can make food or films. Table 1.2 shows how much of each good can be made. This depends on how workers are allocated between the two industries. In each industry, more workers means more output produced. The number of workers is the *scarce resource*. To keep things simple we abstract from monetary values, like the prices of goods and the wages of workers.

Table 1.2 Production possibilities

Food workers	Output	Film workers	Output
4	25	0	0
3	22	1	9
2	17	2	17
1	10	3	24
0	0	4	30

The **law of diminishing marginal returns** applies when one input (such as labour) is varied but other inputs (such as equipment and land) remain fixed. Suppose workers in the film industry can use a fixed number of cameras and studios. The first worker has sole use of these facilities. With more workers, these facilities must be shared. Adding extra workers dilutes equipment per worker. Output per film worker falls as employment rises. A similar story applies in the food industry. Each industry faces diminishing returns to extra workers.

The **law of diminishing marginal returns** says that each extra worker adds less to output than the previous extra worker added.

Table 1.2 shows combinations of food and films made if all workers have jobs. By moving workers from one industry to the other, the economy can make more of one good but only by making less of the other good. There is a trade-off between food output and film output.

The **production possibility frontier (PPF)** shows, for each output of one good, the maximum amount of the other good that can be produced.

Figure 1.4 shows the maximum combinations of food and film output that the economy can produce. Point *A* plots the first row in Table 1.2, where food output is 25 and film output is 0. Points *B*, *C*, *D* and *E* correspond to the other rows of Table 1.2. The curve joining points *A* to *E* in Figure 1.4 is the **production possibility frontier**, or PPF.

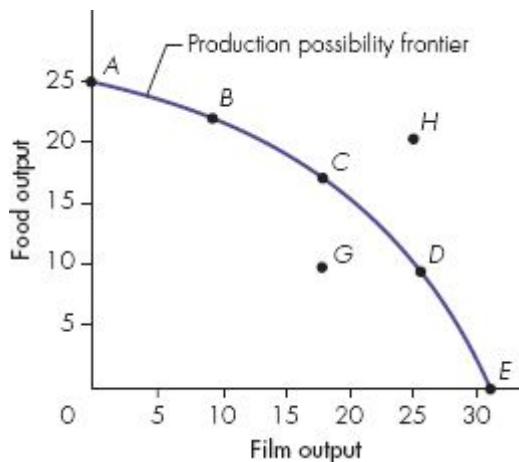


Figure 1.4 The US house price index: monthly data

The frontier shows the maximum combinations of output that the economy can produce using all the available resources. The frontier displays a trade-off: more of one commodity implies less of the other. Points above the frontier need more inputs than the economy has available. Points inside the frontier are inefficient. By fully using available inputs the economy could expand output to the frontier.

The frontier curves around the point given by zero output of both goods. This reflects the law of diminishing marginal returns. Movements from *A* to *B* to *C* each transfer a worker from the food industry to the film

industry. Each transfer yields less additional film output and gives up increasing amounts of food output.

At point *A* we start with 25 units of food but no films. Moving from *A* to *B*, we gain 9 films but lose 3 units of food. Adding an extra worker to the film sector, that is, moving from *B* to *C*, increases the production of films by 8 units but reduces the amount of food by 5 units.

Therefore 3 units of food is the opportunity cost of producing the first 9 films and 5 extra units of food is the opportunity cost of producing 8 extra films. The slope of the PPF tells us the opportunity cost of a good: how much of one good we have to sacrifice to make more of another. Since there are only films and food, the most valued alternative of films is food and *vice versa*.

To see why the curve is a ‘frontier’, think about point *G* in Figure 1.4. Society makes 10 units of food and 17 films. This is feasible. From Table 1.2, it needs 1 worker in the food industry and 2 in the film industry. Society has spare resources. The fourth person is not employed. *G* is not a point on the PPF because we can make more of one good without sacrificing output of the other good. Employing the extra person in the food industry takes us to point *C*, with 7 extra units of food for the same film output. Employing the extra person to work in films takes us to point *D*, with 7 extra units of films but no less food output.

The PPF shows the points at which society is producing efficiently. Points such as *G* inside the frontier are inefficient because society is wasting resources. More output of one good would not require less output of the other. There would be no opportunity cost of expanding output of one good a bit.

Points outside the production possibility frontier, such as *H* in Figure 1.4, are unattainable. Given the inputs available, this output combination cannot be made. Scarce resources limit society to a choice of points inside or on the production possibility frontier.

Since people like food and films, society wants to [produce efficiently](#). Points inside the PPF sacrifice output unnecessarily. Society chooses between the different points *on* the production possibility frontier. In so doing, it decides not only what to produce but how to produce. Table 1.2 shows how many workers must be allocated to each industry to make a particular output combination. As yet, our example is too simple to show for whom society produces.

Production efficiency means more output of one good can be obtained only by sacrificing output of other goods.

How does society decide where to produce on the production possibility frontier? The government may decide. However, in most Western economies, the most important process that determines what, how and for whom goods are produced is the operation of markets.

Opportunity cost and absolute and comparative advantage

The concept of opportunity cost plays a fundamental role in how people make economic decisions. Moreover, the idea of opportunity cost can explain why people trade with each other. Indeed when two individuals (or firms, or nations) have different opportunity costs of performing various tasks, they can always increase the total value of available goods and services by trading with one another. Therefore, the idea of opportunity cost can provide a reason why individuals trade and why trade can be mutually beneficial.

To see how the idea of opportunity cost is related to gains from trade, we consider a simple example. Consider two individuals: Jennifer and John. Both can produce two different goods, cakes and T-shirts. Jennifer is very good at sewing and she can make 4 T-shirts in 1 hour. She is also a good cook and can bake 2 cakes in 1 hour. On the other hand, John can bake 1 cake in 1 hour and he can make a T-shirt in 2 hours. Suppose that both cannot work more than 10 hours a day. Here, time is the scarce resource. The problem is the following: should John and Jennifer produce their own cakes and T-shirts or should they trade with each other?

We are going to look at two possible scenarios.

First scenario: suppose that John and Jennifer produce their own T-shirts and cakes. First, assume that each splits the 10 hours equally in producing the two goods. John can make 2.5 T-shirts and 5 cakes in 10 hours. Jennifer can make 20 T-shirts and 10 cakes. In this scenario the total amount of cakes produced is 15 and the total amount of T-shirts is 22.5.

Second scenario: suppose that John specializes more in producing cakes than T-shirts. For example, assume he spends 8 hours in preparing cakes and 2 hours in making T-shirts. On the other hand, Jennifer specializes more in making T-shirts than cakes. She spends 6 hours in making T-shirts and 4 hours in preparing cakes. In this case, John can make 8 cakes and 1 T-shirt while Jennifer can make 24 T-shirts and 8 cakes. Now the total amount of cakes produced is 16 and we have 25 T-shirts. Compared to the first scenario, the total amount of both goods has increased. Why?

The reason is that, in our example, John and Jennifer have different opportunity costs in producing the two goods. John is more efficient than Jennifer in producing cakes in terms of T-shirts. On the other hand, Jennifer is more efficient in producing T-shirts in terms of cakes than John.

What is John's opportunity cost of cakes in terms of T-shirts? Suppose that John wants to bake more cakes. He needs to spend less time on making T-shirts. How many T-shirts should he forgo? Suppose that John increases the production of cakes by 1 unit. How much time does he need to do that? He needs 1 hour to bake a cake. In that hour he could have made half a T-shirt ($\frac{1}{2}$). So John's opportunity cost of cakes in terms of T-shirts is half a T-shirt.

What about Jennifer? If she wants to bake one more cake she needs an extra half hour. In that half hour she could have made 2 T-shirts. So Jennifer's opportunity cost of cakes in terms of T-shirts is 2 T-shirts.

Therefore, John has a lower opportunity cost of cakes in terms of T-shirts than Jennifer. In this case, we say that John has a **comparative advantage** in making cakes compared to Jennifer. You can check that Jennifer has a comparative advantage in producing T-shirts compared to John.

Therefore, from our example we can say that if each individual specializes more in producing the good in which he or she has a comparative advantage, then it is possible to increase the total production of goods, and so trade can be beneficial.

An individual has a **comparative advantage** compared to another in the production of a good if she has a lower opportunity cost in producing it.

Compared to the first scenario, in the second scenario Jennifer can trade 2 T-shirts with John in exchange for 2 cakes. Jennifer ends up with 22 T-shirts and 10 cakes, while John has 3 T-shirts and 6 cakes. Both gain from this trading in goods compared to the first scenario.

In contrast to the concept of comparative advantage is that of **absolute advantage**. An individual has an absolute advantage in producing a good if he or she is more efficient at producing that good compared to someone else. In our example, Jennifer has an absolute advantage in producing both goods compared to John. In the same period of time she can make more T-shirts and cakes compared to John.

Nevertheless, in determining possible benefits from trade the concept of comparative advantage is what matters, not absolute advantage.

1.4 The role of the market

Markets bring together buyers and sellers of goods and services. In some cases, such as a local fruit stall, buyers and sellers meet physically. In other cases, such as the stock market, business can be transacted by computer. We use a general definition of markets.

In markets buyers and sellers trade goods and services. As a result of this trade process the prices of those goods and services and their quantities (produced and consumed) are determined. Market prices influence the decisions about what, how and for whom to produce.

Prices of goods and of resources (which are also goods, such as labour, machinery, land, energy) adjust to ensure that scarce resources are used to make the goods and services that society wants.

A **market** is a process by which households' decisions about consumption of alternative goods, firms' decisions about what and how to produce, and workers' decisions about how much and for whom to work are all reconciled by adjustment of prices.

The price of a hamburger is lower than the price of a steak. Nevertheless McDonald's is in the business because, given the price of beefburger

meat, the rent and the wages for staff, it can still sell beefburgers at a profit. If rents were higher, it might sell beefburgers in a cheaper area or switch to luxury lunches for rich executives.

The student behind the counter works there because a part-time job helps meet his tuition fees. If the wages were lower, he might not work at all. Conversely, the job is unskilled and there are plenty of students looking for such work, so McDonald's does not have to offer high wages.

ACTIVITY 1.1

SCARCITY IN ACTION: THE RISE OF MARKETS IN A POW CAMP²

In 1945 R. A. Radford, a British economist, published an article about his experience as a prisoner of war during the Second World War. He explained how in a society like the prisoner of war camp economic institutions like markets rose naturally to deal with the problem of scarcity.

The story takes place in an Oflag, a camp for officers only, where between 1200 and 2500 people were imprisoned between 1943 and 1945. There was no paid labour in the camp. The only goods available were the food parcels from the German Army and the Red Cross.

At the beginning the goods rations, such as chocolate, jam and cigarettes, were divided equally among the prisoners. There was no income inequality in that society. However, very soon, prisoners realized that trading could help deal with scarcity. For example, at the beginning non-smokers started giving cigarettes to smokers in exchange for chocolate. Over time more complex exchanges started to take place. A formal shop was created where exchanges could take place. A market was created. Cigarettes became the currency used in exchanges. Each good had a price in terms of cigarettes and only sales in cigarettes were accepted.

Prices changed over time according to demand and supply. Oatmeal, once rare, became more common after 1943 and so its price fell. During hot weather people wanted more soap and less chocolate, and so the price of soap increased while that for chocolate decreased. In 1944 the supply of food parcels was halved

and prices changed accordingly. Canadian butter and marmalade became scarce resources and their prices increased. Prisoners started to switch to German margarine and jam.

Even in a simple society, with no production or labour, a market was created as a natural mechanism to allocate scarce resources.

Questions

- a. How do you think that the creation of the market in the POW camp helped answer the what, how and for whom to produce questions?
- b. If cigarettes become a scarce resource, what do you think should happen to their market price?

Prices guide the consumer's decision to buy a beefburger, McDonald's decision to sell beefburgers and the student's decision to take the job. Society allocates resources – meat, buildings and labour – into beefburger production through the price system. If people hated beefburgers, McDonald's sales revenue would not cover its cost. Society would devote no resources to beefburger production. People's desire to eat beefburgers guides resources into beefburger production.

However, when cattle contract BSE, consumers shun beefburgers in favour of bacon sandwiches, and the price of bacon rises. As the fast-food industry scrambles to get enough pork, the price of pigs rises but the price of beef falls. Adjustments in prices encourage society to reallocate land from beef to pig farming. At the height of the British beef crisis in the mid- 1990s, caused by fears about 'mad cow' disease, pork prices rose 2 per cent but beef prices fell. Quite an incentive to reallocate!

In a **command economy** a government planning office decides what will be produced, how it will be produced and for whom it will be produced. Detailed instructions are then issued to households, firms and workers.

The command economy

How would resources be allocated if markets did not exist? Such planning is very complicated. There is no complete **command economy** where all allocation decisions are undertaken in this way. However, in many countries, for example China, Cuba and those formerly in the Soviet bloc, there was a large measure of central direction and planning. The state owned factories and land, and made the most important decisions about what people should consume, how goods should be produced and how people should work.

This is a huge task. Imagine that you had to run by command the city or town in which you live. Think of the food, clothing and housing allocation decisions you would have to make. How would you decide who should get what and the process by which goods are made and services delivered? These decisions are being made every day, mainly by the allocative mechanism of markets and prices.

CASE 1.3

POOR MARX FOR CENTRAL PLANNERS

During the Cold War, economists used to argue about the relative merits of capitalism and communism. But the Soviet bloc, falling increasingly behind the living standards of the West, abandoned Marxist central planning after 1990 and began transition to a market economy. By 2003 fans of Chelsea Football Club were celebrating the arrival of their new owner, Roman Abramovich, who had made his fortune in the market economy, initially as an oil trader and then as chairman of one of Russia's leading oil companies.

The Berlin Wall fell because the Soviet bloc had fallen far behind market economies in the West. Key difficulties that had emerged were:

- *Information overload* Planners could not keep track of the details of economic activity. Machinery rusted because nobody came to install it after delivery; crops rotted because storage and distribution were not co-ordinated.
- *Bad incentives* Complete job security undermined work incentives. Factory managers ordered excess raw materials to ensure they got materials again the next year. Since planners

could monitor quantity more easily than quality, firms met output targets by skimping on quality. Without environmental standards, firms polluted at will. Central planning led to low-quality goods and an environmental disaster.

- *Insufficient competition* Planners believed big was beautiful. One tractor factory served the Soviets from Latvia in the west to Vladivostok in the east. But large scale deprived planners of information from competing firms, making it hard to assess efficiency. Managers got away with inefficiency. Similarly, without electoral competition, it was impossible to sack governments making economic mistakes.

The ‘invisible hand’

Individuals in **free markets** pursue their own self-interest without government direction or interference. The idea that such a system could solve the what, how and for whom problems is one of the oldest themes in economics, dating back to the Scottish economist Adam Smith, whose book *The Wealth of Nations* (1776) remains a classic. Smith argued that individuals pursuing their self-interest would be led ‘as by an invisible hand’ to do things that are in the interests of society as a whole.

Markets in which governments do not intervene are called **free markets**.

Suppose you wish to become a billionaire. You play around with new ideas and invent something, perhaps a very fuel-efficient car. Although motivated by self-interest, you make society better off by creating new jobs and opportunities. You move society's production possibility frontier outwards – the same resources now make more or better goods – and become a billionaire in the process. Smith argued that the pursuit of self-interest, without any central direction, could produce a coherent society making sensible allocative decisions.

This remarkable insight has been studied at length by modern economists. In later chapters, we explain when the **invisible hand** works

well and when it works badly. Some government intervention may then be justified.

The invisible hand is the assertion that the individual pursuit of self-interest within free markets may allocate resources efficiently from society's viewpoint.

The mixed economy

The free market allows individuals to pursue their self-interest without government restrictions. The command economy allows little scope for individual economic freedom. Decisions are taken centrally by the government. Between these extremes lies the **mixed economy**.

Most countries are mixed economies, though some are close to command economies and others are much nearer to free market economies. Figure 1.5 illustrates this point. Even Cuba allows consumers some choice over the goods they buy. Conversely, even countries such as the US, which espouse more enthusiastically the free market approach, still have substantial levels of government activity in the provision of public goods and services, the redistribution of income through taxes and transfer payments, and the regulation of markets.

In a **mixed economy** the government and private sector jointly solve economic problems. The government influences decisions through taxation, subsidies and provision of free services such as defence and the police. It also regulates the extent to which individuals may pursue their own self-interest.

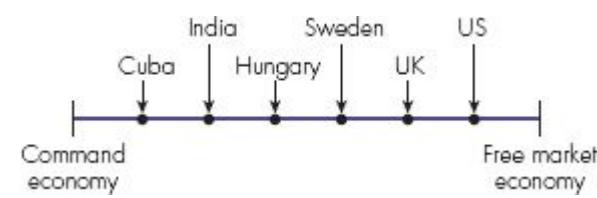


Figure 1.5 Market orientation

The role of the market in allocating resources differs vastly between countries. In the command economy resources are allocated by central government planning. In the free market economy there is virtually no government regulation of the consumption, production and exchange of goods. In between lies the mixed economy, where market forces play a large role but the government intervenes extensively.

1.5

Positive and normative

In studying economics it is important to distinguish between ‘positive’ and ‘normative’ economics.

Positive economics studies objective or scientific explanations of how the economy works.

The aim of **positive economics** is to analyse how society makes decisions about consumption, production and exchange of goods. It aims both to explain why the economy works as it does, and to allow predictions about how the economy will respond to changes. In positive economics, we aim to act as detached scientists. Whatever our political sympathy or our ethical code, we examine how the world actually works. At this stage, there is no scope for personal value judgements. We are concerned with propositions of the form: if *this* is changed, then *that* will happen. In this regard, positive economics is similar to the natural sciences such as physics, geology or astronomy.

Economists of widely differing political persuasions would agree that, when the government imposes a tax on a good, the price of that good will rise. The normative question of whether this price rise is desirable is entirely distinct.

As in any other science, there are unresolved questions where disagreement remains. These disagreements are at the frontier of positive economics. Research in progress will resolve some of these issues but new issues will arise, providing scope for further research.

Normative economics offers recommendations based on personal value judgements.

Competent and comprehensive research can in principle resolve many of the outstanding issues in positive economics; no such claim can be made about the resolution of issues in **normative economics**. Normative economics is based on subjective value judgements, not on the search for any objective truth. The following statement combines positive and normative economics: ‘The elderly have very high medical expenses, and the government should subsidize their health bills.’ The first part of the proposition is a statement in positive economics. It is a statement about how the world works. We can imagine a research programme that could determine whether or not it is correct. (Broadly speaking, it is.)

The second part of the proposition – the recommendation about what the government should do – could never be ‘proved’ true or false by any scientific research investigation. It is a subjective value judgement based on the opinion of the person making the statement. Many people might share this subjective judgement. Others might disagree. You might believe that it is more important to devote society’s scarce resources to improving the environment not the health of the aged.

Economics cannot be used to show that one of these normative judgements is correct and the other is wrong. It all depends on the preferences or priorities of the individual who, or the society that, has to make this choice. But we can use positive economics to clarify the menu of options from which society must eventually make its normative choice.

The professional economist is offering expert advice on positive economics. Scrupulous economists distinguish their role as an expert adviser on positive economics from their status as involved private citizens arguing for particular normative choices.

1.6 Micro and macro

Many economists specialize in a particular branch of the subject. Labour economics deals with jobs and wages. Urban economics deals with land

use, transport, congestion and housing. However, we need not classify branches of economics by subject area. We can also classify branches of economics according to the approach used. In this regard we make a distinction between **microeconomics** and **macroeconomics**.

For example, we can study why individuals prefer cars to bicycles and how producers decide whether to produce cars or bicycles. We can then aggregate the behaviour of all households and all firms to discuss total car purchases and total car production. We can examine the market for cars. Comparing this with the market for bicycles, we can explain the relative price of cars and bicycles and the relative output of these two goods. The sophisticated branch of microeconomics known as general equilibrium theory extends this approach to its logical conclusion. It studies simultaneously every market for every commodity. From this it is hoped to understand the complete pattern of consumption, production and exchange in the whole economy at a point in time.

Microeconomics offers a detailed treatment of how individuals and firms make economic decisions.

Macroeconomics emphasizes interactions in the economy as a whole. It deliberately simplifies the individual building blocks of the analysis in order to retain a manageable analysis of the complete interaction of the economy.

But this is very complicated. It is easy to lose track of the phenomena in which we were interested. The interesting task, which retains an element of art in economic science, is to devise judicious simplifications that keep the analysis manageable without distorting reality too much. Here, microeconomists and macroeconomists proceed down different avenues.

Microeconomists tend to study one aspect of economic behaviour but ignore interactions with the rest of the economy in order to preserve the simplicity of the analysis. A microeconomic analysis of footballers' wages would emphasize the characteristics of footballers and the ability of football clubs to pay. It would largely neglect the chain of indirect effects to which an increase in footballers' wages might give rise (such as higher prices for luxury houses, leading to a boom in swimming pool manufacture). When microeconomic analysis ignores indirectly induced effects, it is 'partial' rather than 'general' analysis.

In some instances, indirect effects may not be important and it will make sense for economists to examine particular industries or activities in great detail. When indirect effects are too important to ignore, an alternative simplification must be found.

Macroeconomists rarely worry about the division of consumer goods into cars, bicycles and DVDs. Instead, they treat them as a single bundle called ‘consumer goods’ because they want to study the interaction between household purchases of consumer goods and firms’ decisions about purchases of machinery and buildings.

Because macroeconomic concepts refer to the whole economy, they get more media coverage than microeconomic concepts, which are chiefly of interest to those in a specific group. Here are three macroeconomic concepts you have probably encountered.

Gross domestic product (GDP)

After the credit crunch in 2007 the global economy entered a recession. Figure 1.6 shows world real **gross domestic product (GDP)** (inflation adjusted). We can see that, after 2007, world real GDP started to decrease compared to previous years. During a recession, GDP is falling or is growing only very slowly.

Gross domestic product (GDP) is the value of total output produced in an economy in a given period.

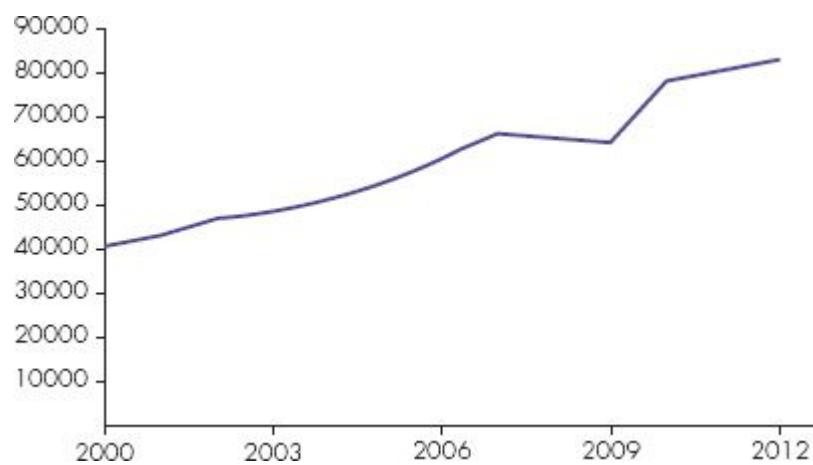


Figure 1.6 World real GDP, US\$ (bn)

Source: © The World Factbook 2013-14. Washington, DC: Central Intelligence Agency, 2013.

Aggregate price level

The prices of different goods may move differently. **aggregate price level** tells us what is happening to prices on average. When this price level is rising, we say there is inflation.

The **aggregate price level** measures the average price of goods and services.

Unemployment rate

The labour force is people of working age who have a job or want one. Some of the rich, the sick and the lazy are of working age but not looking for work. They are not in the labour force and not counted within the **unemployment rate**.

The **unemployment rate** is the fraction of the labour force unemployed but actively looking for a job.

People dislike both inflation and unemployment. In the 1970s, oil price shocks and excessive money creation led to high inflation. Then, inflation fell but unemployment increased. By 2000 both inflation and unemployment had finally fallen back to low levels. Yet by 2007 inflation was beginning to increase again.

However, after the credit crunch took place, in many economies inflation started to fall and unemployment started to rise. In some cases, inflation became negative – a situation known as *deflation*. Macroeconomists want to understand what generates these fluctuations.

Getting the most out of each chapter

There is a summary of the main points at the end of each chapter. Like learning to drive, the best way to check your progress is not to read more and more but to try to do it for yourself. Attempt the review questions that follow the summary (answers are provided on the Online Learning Centre website: www.mcgraw-hill-co.uk/textbooks/begg). Connect has many more problems, also with answers: a self-contained driving instructor.

Summary

- **Economics** analyses what, how and for whom society produces. The key economic problem is to reconcile the conflict between people's virtually unlimited demands and society's limited ability to produce goods and services to fulfill these demands.
- Rational individuals, in making choices, must compare the benefits and the costs associated with those choices. A choice is made only if the benefit of doing it is larger than its cost.
- The **production possibility frontier (PPF)** shows the maximum amount of one good that can be produced given the output of another good. It depicts the trade-off or menu of choices for society in deciding what to produce. Resources are scarce and points outside the frontier are unattainable. It is inefficient to produce within the frontier.
- The **opportunity** cost of an activity is the value of the best alternative that we must sacrifice. It is the slope of the PPF.
- If individuals, firms or countries have different opportunity costs of producing a good compared to others, they have a **comparative advantage**. The fact that individuals have comparative advantages in producing different goods creates the possibility for gains from trading.
- Industrial countries rely extensively on **markets** to allocate resources. The market resolves production and consumption decisions by adjustments in prices.

- In a **command economy**, decisions on what, how and for whom are made in a central planning office. No economy relies entirely on command.
- A **free market economy** has no government intervention. Resources are allocated entirely through markets in which individuals pursue their own self-interest. Adam Smith argued that an ‘invisible hand’ would nevertheless allocate resources efficiently.
- Modern economies are **mixed**, relying mainly on the market but with a large dose of government intervention. The optimal level of intervention is hotly debated.
- **Positive economics** studies how the economy actually behaves. **Normative economics** recommends what should be done. The two should be kept separate. Given sufficient research, economists could agree on issues in positive economics. Normative economics involves subjective value judgements. There is no reason why people should agree about normative statements.
- **Microeconomics** offers a detailed analysis of particular activities in the economy. For simplicity, it may neglect some interactions with the rest of the economy. **Macroeconomics** emphasizes these interactions at the cost of simplifying the individual building blocks.

Review questions



EASY

- 1 Suppose society abolishes higher education and as a result students have to find jobs immediately. If there are no jobs available, how do wages and prices adjust so that those who want jobs can find them?
- 2 Communist Russia used prices to allocate production among different consumers. Central planners set production targets but then put output in shops, fixed prices and gave workers money to spend. Why not plan the allocation of particular goods to particular people as well?
- 3 Which of the following statements are positive and which are normative? (a) Annual inflation is below percent. (b) Because inflation

is low, the government should cut taxes. (c) Income is higher in the UK than in Poland. (d) Brits are happier than Poles.

- 4 **Common Fallacies** Why are these statements wrong? (a) Since some economists are Conservative but others Labour, economics can justify anything. (b) Efficiency gains cannot increase the production of some commodities without sacrificing others, and therefore there is no such thing as a ‘free lunch’. Economics is about people, and thus cannot be a science.
- 5 Which of the following statements refer to microeconomics and which to macroeconomics? (a) Inflation is lower than in the 1980s. (b) The price of a tin of beans fell this month. (c) Good weather means a good harvest. (d) Unemployment in London is below the UK average.

MEDIUM

- 6 OPEC made a fortune for its members by organizing production cutbacks and forcing up prices. (a) Why have coffee producers not managed to do the same? (b) Could UK textile firms force up textile prices by cutting back UK textile production?
- 7 Suppose it becomes possible in 5 years’ time to make as much energy as we want from biofuels provided the price is the equivalent of at least \$50/barrel for oil. (a) What does this imply about the eventual price of oil in, say, 10 years’ time? (b) Is it possible for oil prices to be substantially above \$50/barrel for the next few years? (c) Do higher oil prices in the short run increase or reduce the incentive to look for alternative energy technologies?
- 8 An economy has 5 workers. Each worker can make 4 cakes or 3 shirts. (a) Draw the production possibility frontier. (b) How many cakes can society get if it does without shirts? (c) What points in your diagram are inefficient? (d) Can the economy produce an output combination which lies above the production possibility frontier? (e) What is the opportunity cost of making a shirt and making a cake? (f) Does the law of diminishing returns hold in this economy?
- 9 Suppose that a country can produce two goods: food and clothing. To produce one unit of food, it requires one worker. To produce one unit of clothing, it requires two workers. The total amount of workers available in the economy is fixed and equal to 100. Denoting with L the total amount of workers, with F the units of food produced and with C the units of clothing produced, the resource constraint for this economy can be written as: $L = a_F F + a_C C$, where a_F is the amount of workers

needed to produce one unit of food and α_C the amount of workers needed to produce one unit of clothing. Show how to construct the production possibility frontier from that resource constraint. In a graph with C on the vertical axis and F on the horizontal axis, plot the PPF of this economy. What is the slope of the PPF?

HARD

- |0 Suppose a farmer is planning to grow cabbages on his land. The cost of growing cabbages is £50 per acre and he earns £100 from the produce in the market. There is another option for him, to grow pumpkins, which could yield him £110 if he spent £70 on it. (a) What is the opportunity cost of growing cabbages? Is it rational for the farmer to grow cabbages instead of pumpkins? (c) Suppose the only other option for him to earn from his farmland is to rent it to another farmer. How will the farmer arrive at a rational decision?
- |1 **Essay question** Two similar countries take the decision to try to increase the health of their poorest people. One country raises taxes on the rich and gives more money to the poor. The other country raises taxes on the rich and provides more health care, free to patients, through its national health service. Which country do you think is more likely to meet its objective? Why?

1 Some economists define the opportunity cost as the total cost of an activity, meaning they include the explicit costs in the definition of opportunity cost. We keep them separated to emphasize the fact that the cost of any activity includes both explicit and implicit costs.

2 Based on R. A. Radford, ‘The economic organisation of a P.O.W. camp’, *Economica* 12, no. 48 (1945): 189–201.

CHAPTER 2

Tools of economic analysis

Learning Outcomes

By the end of this chapter, you should be able to:

- 11 recognize why theories deliberately simplify reality
- 12 understand time-series, cross-section and panel data
- 13 construct index numbers
- 14 differentiate between nominal and real variables
- 15 build a simple theoretical model
- 16 understand how to plot data and interpret scatter diagrams
- 17 use ‘other things equal’ to ignore, but not forget, important influences

It is more fun to play tennis if you know how to serve, and cutting trees is much easier with a chainsaw. Every activity or academic discipline has a basic set of tools. Tools may be tangible, like the dentist’s drill, or intangible, like the ability to serve in tennis. This chapter is about the tools economists use. To analyse economic issues we use both *models* and *data*.

Models or **theories** - we use these terms interchangeably – are frameworks to organize how we think about a problem. They simplify by omitting some details of the real world to concentrate on the essentials. From this manageable picture of reality we develop an analysis of how the economy works.

A **model** or **theory** makes assumptions from which it deduces how people will behave. It is a deliberate simplification of reality.

An economist uses a model as a tourist uses a map. A map of Glasgow misses out many features of the real world – traffic lights, roundabouts, speed bumps – but with careful study you get a good idea of how the traffic flows and the best route to take. The simplified picture is easy to follow, but helps you understand actual behaviour when you must drive through the city in the rush hour.

The data or facts interact with models in two ways. First, the data help us quantify the relationships to which our theoretical models draw attention. It is not enough to know that all bridges across the Clyde are likely to be congested. To choose the best route we need to know how long we have to queue at each bridge. We need some facts. The model is useful because it tells us which facts are likely to be the most important.

Second, the **data** help us to test our models. Like all careful scientists, economists must check that their theories square with the *relevant* facts. For example, for a while the number of Scottish dysentery deaths was closely related to UK inflation. Is this a factual coincidence or the key to a theory of inflation? The facts alert us to the need to ponder this question, but we can decide only by logical reasoning.

Data are pieces of evidence about economic behaviour and can be used to test economic models.

In this instance, we can find no theoretical connection. Hence, we view the close factual relationship between Scottish dysentery deaths and UK inflation as a coincidence that should be ignored. Without a logical underpinning, the empirical connection will break down sooner or later. Paying attention to a freak relationship in the data increases neither our understanding of the economy nor our confidence in predicting the future.

The blend of models and data is subtle. The data alert us to logical relationships we had overlooked. And whatever theory we wish to maintain should certainly be checked against the facts. But only theoretical reasoning can guide an intelligent assessment of what evidence has reasonable relevance.

When a theory that makes sense has for a long time survived exposure to the relevant economic data, we sometimes accord it the status of a **behavioural law**, such as the law of diminishing returns.

A **behavioural law** is a sensible theoretical relationship not rejected by evidence over a long period.

Next, we turn to the representation of economic data. Then we show how an economist might develop a theoretical model of an economic relationship. Finally, we discuss how actual data might be used to test the theory that has been developed.

2.1 Economic data

How might we present data to help us think about an economic problem ? There are different ways in which real-world data can be presented. We distinguish between time-series data, cross-section data and panel data.

Time-series data

The first two columns of Table 2.1 report a **time series** of monthly copper prices. It shows how the price changes over time. This information may be presented in tables or charts.

A **time series** is a sequence of measurements of the same variable at different points in time.

Figure 2.1 *plots*, or *graphs*, these data. Each point in the figure corresponds to an entry in the table. Point A shows that in January 2010 the price of copper was \$7385 per tonne. The series of points or dots in Figure 2.1, in whichever colour, contains the same information as the first two columns of Table 2.1.

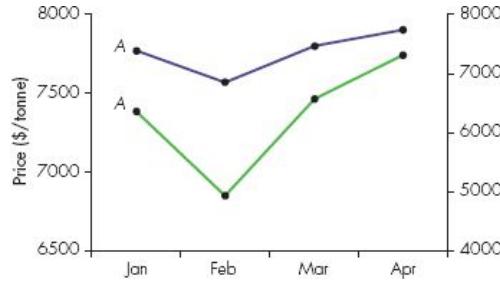
Table 2.1 The price of copper, 2010 (US\$/tonne)

Monthly	\$/tonne	Quarterly	\$/tonne
Jan	7385	IV	6696
Jan	7385	IV	6696
Feb	6847		
Mar	7462		
Apr	7744	I	7231

Source: London Metal Exchange (www.lme.co.uk).

Charts or diagrams must be interpreted with care. The eye is easily misled by simple changes in presentation of the data. In Figure 2.1 the green line corresponds to the left-hand scale and the purple line corresponds to the enlarged scale on the right. Both graphs plot the same data but the blue graph seems to move more. Diagrams can be manipulated in suggestive ways, a point well understood in advertising and politics.

Daily data usually contain too much detail. Imagine studying daily prices over 10 or 20 years!¹ Averages over a month, a quarter (three months) or a year may be the best way to present data. The last two columns of Table 2.1 show quarterly averages for copper prices. The four quarters of the year are the periods January–March, April–June, July–September and October–December. For the fourth quarter of 2009 the quarterly average is \$6696. In the first quarter of 2010 the price of copper was on average \$7231. It can be seen as one-third of the sum of the monthly numbers for January, February and March.



The green line plots Table 2.1 using the left-hand-side scale. The purple line uses the right-hand-side scale. Prices seem now to move less.

Figure 2.1 The monthly price of copper, 2010

Source: London Metal Exchange (www.lme.co.uk)

Cross-section data

Time-series data record how a particular variable changes over time. Economists also use **cross-section data**. Table 2.2 shows a cross-section of unemployment rates in August 2012 for different countries.

Cross-section data record at a point in time the way an economic variable differs across different individuals or groups of individuals.

Table 2.2 Unemployment by country, August 2012 (% of labour force)

US	Japan	Germany	France	UK
8.1	4.2	5.5	10.6	8.0

Source: OECD; ONS.

Panel data

Panel data, also called longitudinal data, are a mix between time-series and cross-section data.

Panel data record observations over multiple time periods for the same individuals or groups of individuals.

Table 2.3 shows a panel data example in which the same variable (unemployment rate) is recorded over time for the same group of countries.

Table 2.3 Unemployment by country, 2008–11 (% of labour force)

	US	Japan	Germany	France	UK
2008	5.7	3.9	7.5	7.2	5.3
2009	9.2	5.0	7.7	9.0	7.7
2010	9.6	5.0	7.0	9.2	7.8

	US	Japan	Germany	France	UK
2011	8.9	4.5	5.9	9.0	7.8

Source: OECD; IMF.

2.2 Index numbers

To compare numbers without emphasizing units of measurement, we use **index numbers**.

An **index number** expresses data relative to a given base value.

Table 2.4 shows annual averages for aluminium and copper prices. We could choose 2004 as the base year and assign the value 100 to both the aluminium and the copper price index in this base year.

Table 2.4 Prices of aluminium and copper (US\$/tonne)

	2004	2007	2010
Aluminium price	1758	2644	2232
Copper price	9.2	6710	7234
Aluminium Index (2004 = 100)	100	150	127
Copper Index (2004 = 100)	100	242	261
Metals Index (2004 = 100)	100	224	234

Source: London Metal Exchange (www.lme.co.uk).

By 2007 the aluminium price of \$2644 per tonne was around 1.5 times its price in 2004. If the aluminium prices had been 100 in 2004, this index must be 150 by 2007. To get the 2010 value, we divide the 2010 aluminium price of \$2232 by the 2004 price of \$1758 to get 1.27. Multiplying this by the starting value of 100 for the index in 2004 yields 127 for the aluminium index in 2010, as in Table 2.4. Using the index number for aluminium we can say that the price of aluminium in 2010 was 27 per cent higher than in 2004, the base year. The price index for copper is calculated in the same way, dividing each price by the 2004 price, then multiplying by 100. According to the copper index, the price of copper in 2010 was 161 per cent higher than in 2004.

Now check that you understand this procedure. In 2001 average aluminium prices were \$1482 per tonne and average copper prices were \$1660. What were the values of the aluminium and copper price indices? (Answer: 84 and 60.)

Index numbers as averages

Now think about the price of metals as a whole. The prices of different metals change differently. To derive a single measure of metal prices we *average* different metal prices.

Suppose aluminium and copper are the only metals. An index of metal prices in the fifth row of Table 2.4 makes a single time series by combining the time series in the third and fourth rows. In the metal index, each metal has a weight or share that reflects the purpose for which the index is constructed. If it summarizes what firms pay for metal inputs, the weights should reflect the relative use of aluminium and copper as industrial inputs. Copper is much more widely used than aluminium. We might choose a weight of 0.8 for copper and 0.2 for aluminium. The weights always add up to 1.

The last row of Table 2.4 shows changes over time in the metal price index, the *weighted average* of the indices for aluminium and copper. In the base year 2004, the metals index is 100, being $(0.2 \times 100) + (0.8 \times 100)$. By 2007 the index is around 224, which is $(0.2 \times 150) + (0.8 \times 242)$. In 2010 the index was 234.

The metals index, a weighted average of aluminium and copper prices, must lie between the indices for the two separate metals. The weights determine whether the metals index more closely resembles

the behaviour of copper prices or aluminium prices.

The CPI and other indices

To keep track of the prices faced by consumers, countries construct a [consumer price index \(CPI\)](#). The CPI is used to measure changes in the cost of living; that is, the money that must be spent to purchase the typical bundle of goods consumed by a representative household. In the UK the CPI forms the basis for the government's inflation target which the Bank of England's Monetary Policy Committee is required to achieve. The CPI is constructed in two stages. First, index numbers are calculated for each category of commodity purchased by households. Second, the CPI is constructed by taking a weighted average of the different commodity groupings. Table 2.5 shows the weights used and the main commodity groupings. The weights sum up to 1. The 'shopping basket' described in Table 2.5 is reviewed every year to make sure that it is up to date and representative of consumers' spending.

The [consumer price index \(CPI\)](#) measures changes in the cost of living by looking at the cost of a standard 'shopping basket' of goods.

Table 2.5 Prices of aluminium and copper (US\$/tonne)

Item	Weights
Food and non-alcoholic beverages	0.106
Alcoholic beverages and tobacco	0.044
Clothing and footwear	0.068
Housing and household services	0.137
Furniture and household goods	0.059
Health	0.025
Transport	0.148
Communication	0.31
Recreation and culture	0.141
Education	0.021
Restaurants and hotels	0.117
Miscellaneous goods and services	0.103

Source: ONS (2013) Consumer Prices Index and Retail Prices Index: The 2013 Basket of Goods and Services.

A 10 per cent rise in food prices will change the CPI more than a 10 per cent rise in the price of alcoholic beverages and tobacco. This is because food has a much larger weight than alcohol and tobacco in consumers' expenditure.

Another price index that is particularly important is the *retail price index (RPI)*, which is also used to measure changes in the cost of living. The RPI is similar to the CPI, the main differences being in terms of the items included in one index and not in the other. For example, the RPI includes mortgage interest payments by household while the CPI does not. In the past, the RPI used to be the index used to calculate inflation in the UK. More recently, the CPI has been adopted as the main measure for UK inflation, as it is in other European countries. Price indices like the CPI and RPI may overstate the cost of living because they do not measure accurately changes in the quality of the goods over time.

In Figure 2.2 we plot the [inflation rate](#) in the UK for the period 1989-2012 as measured by CPI and RPI indices. While there are differences between the two inflation measures, the behaviour of inflation over time looks pretty similar in both indices.

The [inflation rate](#) is the annual rate of change of the consumer price index.

Other examples of indices include the index of wages in manufacturing, a weighted average of wages in different manufacturing industries. The FTSE, or ‘footsie’, is the *Financial Times-Stock Exchange* index of share prices quoted on the London Stock Exchange. The *index of industrial production* is a weighted average of the *quantity* of goods produced by industry.

The process by which index numbers are calculated is always the same. We choose a base date at which to set the index equal to 100, then calculate other values relative to this baseline. Where the index refers to more than one commodity, we have to choose weights by which to average across the different commodities that the index describes.

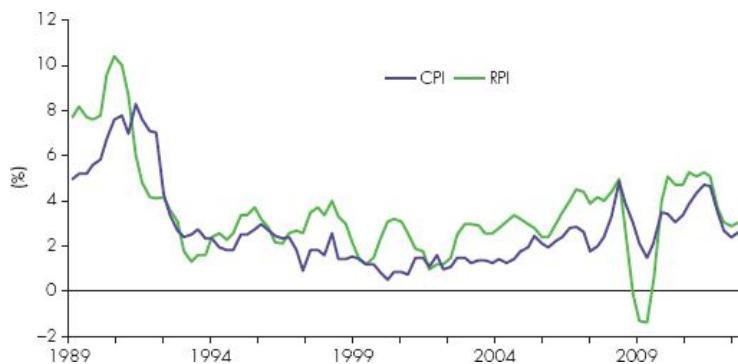


Figure 2.2 UK inflation rate

Source: Office for National Statistics.

CONCEPT 2.1

HYPERNFLATION

In developed countries over the last few decades inflation has been moderate. However, there are cases in which inflation explodes over time. When prices start to increase very rapidly over time and inflation becomes particularly large, we have a case of hyperinflation. A typical example is what happened in Germany after the First World War. In 1918 the Allied victors demanded that Germany make reparations for the damage done and pay the pensions of Allied armed forces engaged in the war. By 1922, in economic ruin, Germany suspended reparations. In January 1923 French and Belgian troops occupied the Ruhr coalfields. German workers began a general strike and the government rolled the presses to print money to pay the 2 million workers involved.

This was the last straw for the German economy. Prices spiralled out of control. Monthly inflation reached the equivalent of 1 million per cent a year. Paper money became almost worthless.

Hyperinflation is not particularly rare. In the 1980s it occurred in several Latin American countries (Bolivia, Argentina, Brazil). More recently, Zimbabwe experienced extreme hyperinflation. To give an idea of the magnitude of price level increases, in 2008 the inflation rate in Zimbabwe was 231 150 888.87 per cent. This means that if a good had a price of 1 Zimbabwean dollar in 2007, the same good would have a price of 231 150 888.87 Zimbabwean dollars in 2008! During hyperinflation, national currency becomes worthless and people simply do not want to use it. In 2009 Zimbabwe abandoned printing of the Zimbabwean dollar, and the South African rand and US dollar became the standard currencies for exchange.

2.3

Nominal and real variables

The first row of Table 2.6 shows the average price of a house, which rose from £2800 in 1963 to £166 000 in 2012.² Are houses really 53 times as expensive as in 1963? Not when we allow for inflation, which also raised incomes and the ability to buy houses.

Table 2.6 Average UK house prices

	1963	1983	2012
House price (£000s)	2.8	27.6	166
RPI (2012 = 100)	4.5	34.3	100
Real price of houses (2012 £000s)	62.2	80.4	166

Source: Nationwide.

The second row of Table 2.6 shows the retail price index, using 2012 as the base year.³ Inflation led to substantial increases in the RPI during 1963–2009. The third row of Table 2.6 calculates an index of real house prices, expressed in 2012 prices. The value of house prices is the same in 2012 in the top and bottom rows.

To calculate the real price of houses in 1963, by expressing them at 2012 prices, we take the nominal price of £2800 and multiply by $[(100)/(4.5)]$ to allow for subsequent inflation, yielding £62 200. Real prices have roughly increased by a factor of 2.5 since 1963 (from £62 200 to £166 000). Most of the 53-fold increase in nominal house prices in the top row of Table 2.5 was due to inflation.

Real or relative prices

The distinction between **nominal** and **real values** applies to all variables measured in money values. It does not apply to units of output, such as 4000 carpets per annum, which relate to physical quantities. Whatever the inflation rate, 4000 carpets is 4000 carpets. However, we do not know whether £100 is a large or a small amount until we know the measurement. general price level for goods.

Nominal values are measured in the prices ruling at the time of measurement.

Real values adjust nominal values for changes in the price level.

The argument carries over to prices themselves. The nominal price of silver has risen a lot since 1970. To calculate an index of the *real price of silver*, divide an index of nominal silver prices by the RPI or the CPI and multiply by 100. Real prices indicate economic scarcity. They show whether the price of a commodity rose faster than prices in general. Hence, real prices are sometimes called *relative prices*.

Consider the price of televisions over the past 20 years. TV prices, measured in pounds, have hardly changed. The RPI and the CPI have risen. The real price of TVs has fallen. Advances in technology have increased the quality of televisions but have also reduced the cost of producing them. Because the real price has fallen, many households now have several TVs. It is misleading to base our analysis on nominal values of variables.

The purchasing power of money

When the price of goods rises, the **purchasing power of money** falls because £1 buys fewer goods. To distinguish between real and nominal variables, we say that real variables measure nominal variables as if the purchasing power of money had been constant. Another way to express this idea is to distinguish nominal variables in *current* pounds and real variables in *constant* pounds.

The **purchasing power of money** is an index of the quantity of goods that can be bought for £1.

Table 2.6 described real prices of houses measured in 2012 pounds. We could of course have used 1960 pounds instead. Although the level of the real price index for houses would have been different, it would have grown at exactly the same rate as in the final row of Table 2.6.

CONCEPT 2.2

MONEY ILLUSION

The distinction between nominal and real variables is a key concept in economics. Money illusion refers to a tendency to think in terms of nominal rather than real monetary values. This means that the nominal value of money can be mistaken for its purchasing power. Suppose you work and you get a wage of £1000. With that wage you buy only bread and the price of bread is £1 per kg. This means your wage in real terms has a value of 1000 kg of bread (meaning that the purchasing power of your wage is 1000 kg of bread).

Now suppose you are asked to choose between the following two cases:

- [1] You can get an increase to £1600 in your wage while the price of bread is £2.
- [2] You can get a reduction in your wage to £800 while the price of bread is £1.

Which one should you choose ? The two situations are equivalent in ‘real terms’. Therefore if you are rational you should not prefer one to the other. In both cases you buy the same amount of bread (800 kg). People who perceive these situations differently are said to be prone to money illusion.

Do people suffer from money illusion ? Some evidence suggests that they do. Indeed, if we asked many individuals the same question as above, we would probably see some individuals choosing case (1). The reason is that some individuals will think that an increase in the nominal wage is better than a decrease in the nominal wage, even if in real terms nothing has changed.

Source: E. Shafir et al., ‘On money Illusion’, *Quarterly Journal of Economics* 112, no. 2 (1997): 341-374.

2.4 Measuring changes in economic variables

During the BSE crisis in 1996, UK beef production fell from 90 000 tonnes in January to 50 000 tonnes in April. The *absolute change* was $-40\ 000$. The minus sign tells us it fell. The *percentage change* in UK beef output was $(100) \times (-40\ 000)/(90\ 000) = -44\%$. Absolute changes specify units (e.g. tonnes), but percentage changes are *unit-free*. Data are often shown this way.

The *percentage change* is the absolute change divided by the original number, then multiplied by 100.

The *growth rate* is the percentage change per period (usually a year).

When we study time-series data over long periods such as a decade, we do not want to know just the percentage or absolute change between the initial date and the final date. We want to know the percentage or absolute change between every period. Negative growth rates show percentage falls. Economists usually take *economic growth* to mean the percentage annual change in the national income (that is, the real gross domestic product).

2.5 Economic models

Now for an example of economics in action. The London Underground, known locally as the tube, usually loses money and needs government subsidies. Might different policies help ? You have to set the tube fare that will raise most revenue. How do you analyse the problem ?

To organize our thinking, or build a model, we need to simplify reality, picking out the key elements of the problem. Economic models are usually built using mathematical equations linking the variables of interest. We begin with the simple equation

$$\text{Revenue} = [\text{fare}] \times [\text{number of passengers}] \quad (1)$$

London Underground sets the fare and influences the number of passengers only through the fare that is set. (Cleaner stations and better service may help. We neglect these for the moment.)

The number of passengers may reflect habit, convenience and tradition, and be completely unresponsive to changes in fares. This is *not* the view an economist would adopt. It is possible to travel by car, bus, taxi or tube.

Decisions about how to travel will depend on the relative costs of different modes of transport. Equation (1) requires a ‘theory’ or ‘model’ of what determines the number of passengers. We must model the *demand* for tube journeys.

First, the tube fare matters. Other things equal, higher tube fares reduce the number of tube journeys demanded. Second, if there are price rises for the competing modes of taxis and buses, more people will use the tube at any given tube fare. Third, if passengers have higher income, they can afford more tube journeys at any given fare. We now have a bare-bones model of the number of tube passengers:

$$\text{Number of passengers} = f(\text{tube fare}, \text{taxi fare}, \text{petrol price}, \text{bus fare}, \text{passenger incomes} \dots) \quad (2)$$

The number of passengers ‘depends on’ or ‘is a function of’ the tube fare, the taxi fare, petrol prices, bus fares, incomes and some other things. The notation $f(\dots)$ is shorthand for ‘depends on all the things listed inside the brackets’; or, in mathematical terms, it reads that the number of passengers is a function of all the variables in the brackets. The row of dots reminds us that we have omitted some possible determinants of demand to simplify our analysis. Tube demand probably depends on the weather. It is uncomfortable in the tube when it is hot. If the purpose of our model is to study *annual changes* in the number of tube passengers, we can neglect the weather provided weather conditions are broadly the same every year.

Writing down a model forces us to look for all the relevant effects, to worry about which effects must be taken into account and which can be ignored in answering the question we have set ourselves. Combining equations (1) and (2),

$$\begin{aligned} \text{Tube revenue} &= \text{tube fare} \times \text{number of passengers} \\ &= \text{tube fare} \times f(\text{tube fare}, \text{taxi fare}, \text{petrol price}, \text{bus fare}, \text{incomes} \dots) \end{aligned} \quad (3)$$

Why all the fuss? You would have organized your approach along similar lines. That is the right reaction. Models are simply devices to ensure we think clearly about a problem. Clear thinking requires simplification. The real world is too complicated for us to think about everything at once. Learning to use models is more an art than a science. Too much simplicity will omit a crucial factor from the analysis. Too much complexity and we lose any feeling for why the answer turns out as it does.

Sometimes data guide us about which factors are crucial and which are not. At other times, as with tube fares, it is not enough to understand the forces at work. We need to quantify them. For both reasons, we turn now to the interaction of economic models and economic data.

2.6 Models and data

Equation (3) is our model of determinants of tube revenue. Higher fares give *more* revenue per passenger, but *reduce* the number of passengers. Theory cannot tell us which effect dominates. This is an *empirical* or factual issue: how many passengers are put off by higher fares?

Empirical evidence

We need some empirical research to establish the facts. *Experimental* sciences, including many branches of physics and chemistry, conduct controlled experiments in a laboratory, varying one factor at a time while holding constant all the other relevant factors. Like astronomy, economics is primarily a *non-experimental* science. Astronomers cannot suspend planetary motion to examine the relation between the earth and the sun in isolation; economists cannot suspend the laws of economic activity to conduct controlled experiments.

Most empirical research in economics must deal with data collected over periods in which many of the relevant factors were simultaneously changing. The problem is how to disentangle the separate influences on observed behaviour. We approach this in two stages. First, we proceed by examining the relationship of interest - the dependence of revenue on fares - neglecting the possibility that other relevant factors were changing. Then we indicate how economists deal with the harder problem in which variations in other factors are also included in the analysis.

Table 2.7 shows data on tube fares and passengers. When annual data are measured over overlapping calendar years - say, from April 1999 to March 2000 - we show the year as 1999/00. Column (1) shows the real tube fare per passenger kilometre, column (2) shows tube demand, in billions of passenger kilometres a year, and column (3) shows real revenue.

Table 2.7 The tube, 1999/00–2008/09

	(1) Real fare (08/09 pence)	(2) No. of trips (bn pass. km)	(3) Real revenue (08/09 £m)
1999/00	18.4	7171	1319
2000/01	18.6	7470	1389
2001/02	18.6	7451	1386
2001/03	18.0	7367	1326
2001/04	17.9	7340	1314
2001/05	18.1	7606	1377
2001/06	18.7	7586	1419
2001/07	18.8	7947	1494
2001/08	18.7	8352	1562
2001/09	18.7	8646	1617

Source: Department of Transport (www.dft.gov.uk). © Crown copyright 2010.

It is useful to present evidence such as that in Table 2.7 in a [scatter diagram](#), such as that shown in Figure 2.3. The horizontal axis measures column (3), the real fare per passenger kilometre. The vertical axis measures column (1), real revenue in constant million pounds. Real revenue is the real fare per passenger kilometre multiplied by the number of passenger kilometres travelled. Each point in the graph denotes the value of real revenues and the real fare in a particular period. For example, the bottom dot on the left denotes the year 1999/00, where the real fare was 18.4 and the real revenues were 1319.

A [scatter diagram](#) plots pairs of values simultaneously observed for two different variables.

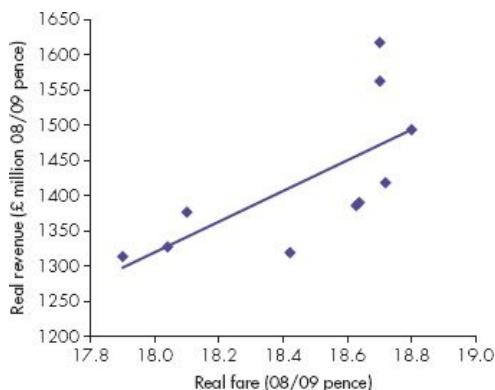


Figure 2.3 Tube fares and revenues, 1999/00–2008/09

Source: Department of Transport (www.dft.gov.uk) © Crown copyright 2010.

From Figure 2.3 we can see a positive relationship between real fare and real revenue. Other things equal, higher fares reduce the number of tube journeys, but if quantity demanded falls only a little, overall revenue may rise when fares are increased. Certainly, in some years, passenger use rose strongly despite higher fares. But we have not yet got to the bottom of things. We return to this issue in Section 2.8.

2.7 Diagrams, lines and equations

If we can draw a line or curve through all these points, this suggests, but does not prove, an underlying relationship between the two variables. If, when the points are plotted, they lie all over the place, this suggests, but does not prove, no underlying relationship between the two variables. Only if economics were an experimental science, in which we could conduct controlled experiments guaranteeing that all other relevant factors had been held constant, could we interpret scatter diagrams unambiguously. Nevertheless, they often provide helpful clues.

Fitting lines through scatter diagrams

In Figure 2.3 we did draw a line through the scatter of points we plotted. The line shows the average relation between fares and revenue between 1999/00 and 2008/09. We can quantify the average relation between fares and usage.

Given a particular scatter of points, how do we decide where to draw the line, given that it cannot fit all the points exactly? The details need not concern us here, but the idea is simple. Having plotted the points describing the data, a computer works out where to draw the line to minimize the dispersion of points around the line.

After some practice, most people can work with two-dimensional diagrams such as Figure 2.3. A few gifted souls can even draw diagrams in three dimensions. Fortunately, computers can work in 10 or 20 dimensions at once, even though we cannot imagine what this looks like.

This solves the problem of trying to hold other things constant. The computer measures the tube fare on one axis, the bus fare on another, petrol prices on a third, passenger incomes on a fourth and tube revenue on a fifth, plots all these variables at the same time, and fits the average relation between tube revenue and each influence when they are simultaneously considered. Conceptually, it is simply an extension of fitting lines through scatter diagrams.

By disentangling separate influences from data where many different things move simultaneously, econometricians conduct empirical research to test economic models.

Reading diagrams

You need to be able to read a diagram and understand what it says. Figure 2.4 shows a hypothetical relationship between two variables: P for price and Q for quantity. The diagram plots $Q = f(P)$. This notation means that the variable Q is related to the variable P through the function f . If we know the function f , knowing the value of P tells us the corresponding value of Q . We need to know values of P to make statements about Q . In Figure 2.4, Q is a *positive* function of P . Higher values of P imply higher values of Q .

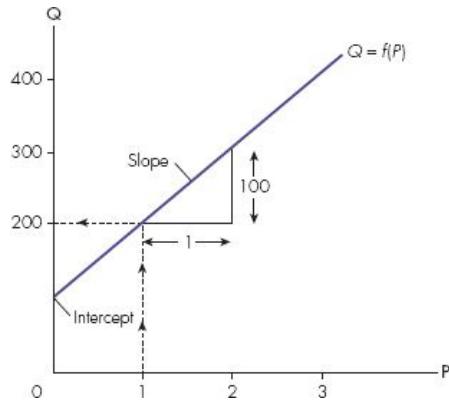


Figure 2.4 A positive linear relationship

When, as in Figure 2.4, the function is a straight line, only two pieces of information are needed to draw in the entire relationship between Q and P . We need the *intercept* and the *slope*. The intercept is the height of the line when the variable on the horizontal axis is zero. In Figure 2.4, the intercept is 100, the value of Q when $P = 0$.

Lots of different lines could pass through the point at which $Q = 100$ and $P = 0$. The other characteristic is the *slope* of the line, measuring its steepness. The slope tells us how much Q (the variable on the vertical axis) changes each time we increase P (the variable on the horizontal axis) by one unit. In Figure 2.4, the slope is 100. By definition, a straight line has a constant slope. Q rises by 100 whether we move from a price of 1 to 2, or from 2 to 3, or from 3 to 4. The equation of the straight line plotted in Figure 2.4 is

$$Q = 100 + 100P$$

Therefore in this case we have: $f(P) = 100 + 100P$.

Figure 2.4 shows a *positive* relation between Q and P . Since higher P values are associated with higher Q values, the line slopes *up* as we increase P and moves to the right. The line has a positive slope. Figure 2.5 shows a case where Q depends *negatively* on P . Higher P values now imply smaller Q values. The line has a negative slope.

The equation of the straight line plotted in Figure 2.5 is

$$Q = 300 - 100P$$

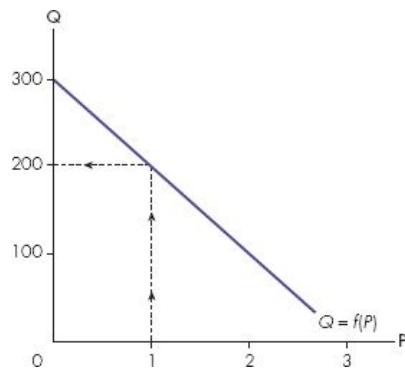


Figure 2.5 A negative linear relationship

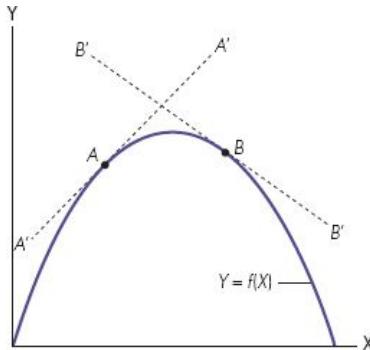


Figure 2.6 A non-linear relationship

Economic relationships need not be straight lines or linear relationships. Often relationships between economic variables are non-linear. Figure 2.6 shows a non-linear relationship between two variables, Y and X . The slope keeps changing. Each time we raise X by one unit we get a different rise (or fall) in Y . Consider the relationship between the income tax rate X and income tax revenue Y . When the tax rate is zero, no revenue is raised. When the tax rate is 100 per cent, nobody bothers to work and revenue is again zero. Beginning from a zero tax rate, rises in tax rates initially raise total tax revenue. Beyond some tax rate, further rises in tax rates then reduce tax revenue, which becomes zero by the time the tax rate is 100 per cent. Diagrams display the essence of real-life problems.

An equation that can give rise to the graph in Figure 2.6 is

$$Y = 100x - x^2$$

The equation above is non-linear since, on the right-hand side, there is a variable with an exponent different from 1.

ACTIVITY 2.1

LANDING THE BIG JOB

Two students, David and Samantha, have to decide how hard to work for the final exam.

They need a mark of 70 to get the job with Greenpeace that they want. Their tutor has promised them that the exam will be just as hard (or easy) as previous exams. David and Samantha have all the marks from their previous exams, and also know how hard they worked (minimum effort is 1, maximum effort is 5, and neither student worked all that hard in the exams leading up to their

finals). From past experience, they know there is a linear relationship between effort and exam results.

Exam marks	Effort level		
	1	2	3
David	20	40	60
Samantha	30	60	90

Questions

- a. What effort level must David make in order to land his job with Greenpeace ?
- b. What effort level does Samantha have to make ?
- c. Which student is better at exams ?
- d. Give three possible reasons for the different exam performances of David and Samantha.

To check your answers to these questions, go to page 677.

2.8 Another look at ‘other things equal’

A diagram might help London Underground think about tube fares. Apart from tube fares, the key determinants of passenger use are probably the incomes that passengers have available to spend, and the introduction of the congestion charge in 2003, which induced some Londoners to abandon their cars in favour of public transport.

In the period 1999/00-2008/09, Britain’s national income, adjusted for inflation, grew substantially, with the exception of the year 2009 because of the economic recession. Look again at Table 2.7. Even if tube fares had been constant, rising incomes should have led to (and did lead to) rising tube use and rising tube revenues.

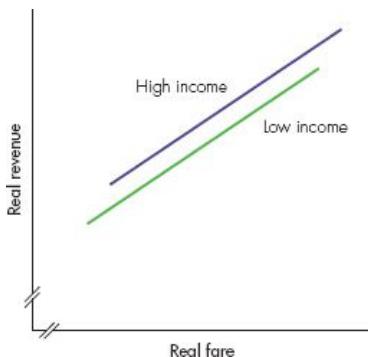


Figure 2.7 Other things equal

Once we allow for movements in *both* tube fares *and* incomes of passengers, our analysis makes more sense. Imagine two sub-periods, one in which incomes were low and one in which incomes were high. Figure 2.7 shows the relationship between tube fare and tube revenue in each period separately. The green line corresponds to low incomes and hence low passenger demand for tube journeys. The purple line shows greater demand for the tube at each and every potential level of tube fares.

During 1999/00-2008/09 we moved from points near the bottom of the purple line to points near the top of the green line. Tube revenue increased not merely because fares rose but also because incomes rose. Similarly, the introduction of the congestion charge for car use in London increased the demand for tube journeys at any particular level of fares and income.

The axes of Figure 2.7 encourage us to think about the relationship between fares and revenue. [Other things equal](#), higher fares yield higher revenue and we move *along* the upward-sloping line. When one of these other things (like income or congestion charges) changes, we show this as a *shift* in the line. Now we can draw two-dimensional diagrams without neglecting other determinants. When things not drawn on the axes change, lines (or curves) shift.

[Other things equal](#) is a device for looking at the relationship between two variables, but remembering that other variables also matter.

The same reasoning applies to the introduction of the congestion charge on vehicle use after 2003. Other things equal, fewer people drove cars and more people used the tube. As with the rise in income in Figure 2.7, the consequence of greater tube use is to generate more revenue at any particular level of tube fare. That is why, in Table 2.7, the data after 2003/04 show much higher tube use than before at each level of tube fare.

2.9 Theories and evidence

Economists analyse a problem in three distinct stages. First, a phenomenon is observed or contemplated and the problem is formulated. By armchair reasoning or a cursory inspection of the data, we decide tube fares have something to do with tube revenues. We want to understand what this relationship is and why it exists.

Second, we develop a theory or model to capture the essence of the phenomenon. By thinking about incomes and the decision about which type of transport to use, we identify the things relevant to tube usage and hence tube revenue.

Third, we test the predictions of the theory by confronting it with economic data. An [econometric](#) examination of the data can quantify the things the model identifies. We can see if, on average, they work in the direction our model suggests. Indeed, by including in our econometric investigation some extra factors deliberately left out of our model in the quest for simplicity, we can check that the extra influences were sufficiently unimportant that it made sense to omit them from the analysis.

[Econometrics](#) is the term used to refer to statistics used to measure relationships in economic data.

Suppose we confront our theory with the data and the two seem compatible. We *do not reject* our theory. If our model is rejected, we have to start again. If our model is not rejected by the data, this does not guarantee that we have found the correct model. There may be a better model that has escaped our attention but would also be compatible with our particular data. As time passes, we acquire new data. We can also use data from other countries. The more we confront our model with different data and find that it is still not rejected, the more confident we become that we have found the true explanation of the behaviour in which we are interested. Relationships in which we have become very confident are sometimes called *economic laws*.

2.10 Some popular criticisms of economics and economists

This chapter has introduced the economist's toolkit. You may have some nagging doubts about it. We end the chapter by discussing some of the popular criticisms of economics and economists.

CASE 2.1

GET A BECKER VIEW: USE AN ECONOMIST'S
SPECTACLES

Most people accept that the economic analysis of markets - thinking about how incentives affect resource allocation - helps us understand things like inflation or unemployment. Can the same tools be applied to other social behaviour; to crime; to marriage; to drug use?

Since much of economic analysis supposes that people are driven by self-interest, rather than by an altruistic concern for others, some economists doubt whether economics can shed light on highly interactive 'social' situations. Other economists have no such fears. In 1992 Chicago economist Gary Becker was awarded the Nobel Prize for Economics for applying the logic of economic incentives to almost every facet of human behaviour. Some examples of Becker in action ...

Marriage and divorce

'The courtroom is not a good place to make judgments about the unique circumstances of each marriage or relationship. We should replace judicial determination with marriage contracts that specify, among other things, the financial and child custodial terms of a divorce. Marriage contracts would become much more common if we set aside the legal tradition that they are unenforceable.'



Gary Becker © www.nutquote.com

Drugs

Prohibition of alcohol gave the US Al Capone but failed to stop drinking. The end of Prohibition 'was a confession that the US experiment in banning drinking had failed dismally. It was not an expression of support for heavy drinking or alcoholism.' Becker's solution for drugs is to legalize, boost government tax revenue, protect minors and cut out organized crime's monopoly on supply.

Becker's proposals have some merit. For example, in 2001 the police in Brixton were told not to arrest people smoking cannabis in public, part of a gradual policy switch to target hard drugs like cocaine and heroin. UK cannabis seizures in 2000 were only half those of 1997. As supply increased, the price on the street slumped.

Some people argue that implicit toleration of soft drugs should give way to decriminalization, allowing legal sales. With 1500 tonnes consumed annually in the UK, an excise duty of £3 a gramme would raise up to £5 billion a year in tax revenue. Gains to the wider economy would be even greater. There would probably be cuts both in the £1.4 billion currently spent enforcing anti-drugs laws and in the £1.5 billion estimated as the cost of drug-related crime.

Source: G. S. Becker and G. N. Becker, *The economics of life* (McGraw-Hill, 1997); *The Observer*, 8 July 2001.

No two economists ever agree

You need to distinguish between positive economics and normative economics. Even if all economists agree on the positive economics of how the world works, there is huge scope to disagree on normative recommendations based on different value judgements. Many disagreements between economists fall under this heading.

There *are* disagreements in positive economics. Economics is only rarely an experimental science. It is prohibitively costly to make half of the population unemployed just to find out how the economy then

works. Without controlled experiments, we have to disentangle different influences in past data to overcome the problem of other things equal. Using data over many years makes it easier to do this unravelling but introduces a new problem. Since attitudes and institutions are slowly changing, data from many years ago may no longer be relevant to current behaviour. The problems we confront are difficult ones and we have to do the best we can.

Finally, it is wrong to think that there are not serious disagreements between physicists or doctors or engineers. Most people do not pretend to know much about physics; everybody claims to know a bit about the problems that economists study.

Models in economics are so simple they have little to do with reality

A model is a deliberate simplification to help us think more clearly. A good model simplifies a lot but does not distort reality too much. It captures the main features of the problem. The test of a good model is not how simple it is, but how much of observed behaviour it can explain.

Sometimes we can get a long way with a simple model. You will see examples in later chapters. On other occasions, the behaviour we are studying is complex and a simple model may not suffice. Where a more realistic model would take us beyond the scope of this book, we still introduce a simple model to let you begin to see the elements of the problem.

People are not as mercenary as economists think

Economists believe that most of the phenomena they study, such as whether to travel by bus or by tube, are mainly determined by economic incentives. This does not mean economic incentives are all that matter.

A successful advertising campaign by the tube would change tube usage. So would a change in social attitudes: it might become chic to take the tube. Knowledge of politics, sociology and psychology is needed for a more complete description of human behaviour. These are factors that economists subsume under the heading of ‘other things equal’. Economics emphasizes the effect of economic incentives. Social attitudes change slowly and for many purposes may be treated as being held constant. However, if an economist discovered an important change in social attitudes, it would be easy to include this in the analysis.

Actions of human beings cannot be reduced to scientific laws

Physicists accept that individual molecules behave randomly but that we can construct and test theories based on their average or systematic behaviour. Economists take the same view about people. We shall never explain actions based on whim or because you got out of bed on the wrong side. However, random differences in behaviour tend to cancel out, on average. We can describe average behaviour with a lot more certainty.

If behaviour shows no systematic tendencies - tendencies to do the same thing when confronted by the same situation - there is little to discuss. The past is no guide to the future. Every decision is a one-off decision. Not only is this view unhelpful, it is not supported by the data. The economic theories that survive are those consistently compatible with the data. The more random is human behaviour, the smaller is the systematic element about which we can form theories and use to make predictions. It is better to be able to say something about behaviour than nothing at all. Often, as you will shortly discover, we can say rather a lot.

Summary

- There is a continuing interplay between models and data in the study of economic relationships. A **model** is a simplified framework to organize how we think about a problem.

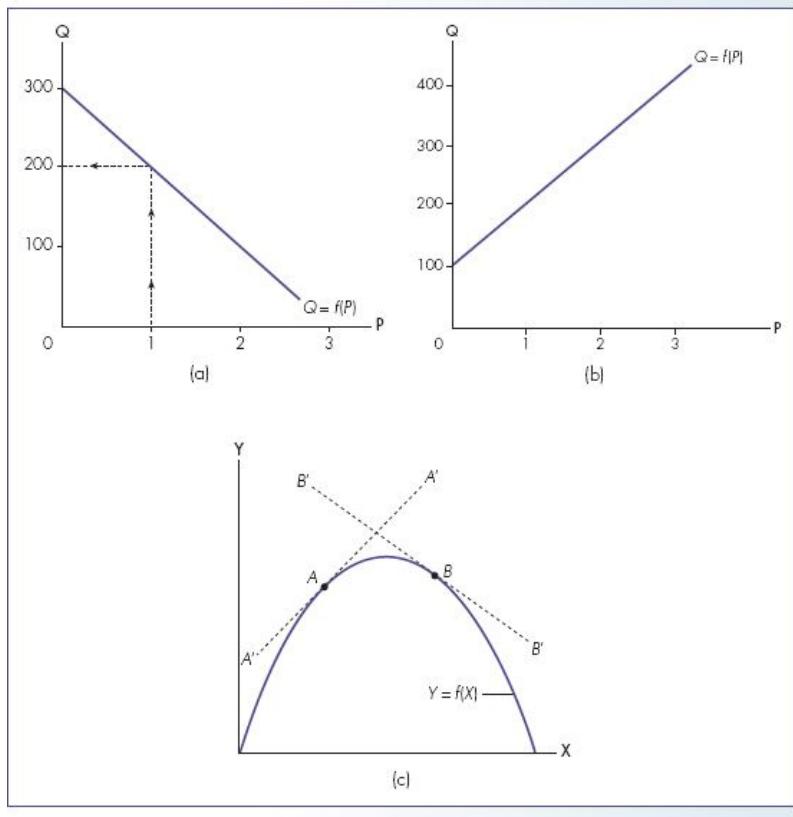
- **Data** or facts are essential for two reasons. They suggest relationships which we should aim to explain and they allow us to test our hypotheses and to quantify the effects that they imply.
 - Tables present data in a form easily understood. **Time-series data** are values of a given variable at different points in time. **Cross-section data** refer to the same point in time but to different values of the same variable across different people. **Panel data** are a mix between time-series and cross-section data.
 - **Index numbers** express data relative to some given base value.
 - Many index numbers refer to averages of many variables. The **consumer price index** summarizes changes in the prices of all goods bought by households. It weights the price of each good by its importance in the budget of a typical household.
 - The annual percentage change in the retail price index is the usual measure of **inflation**, the rate at which prices in general are changing.
 - **Nominal or current price variables** refer to values at the prices ruling when the variable was measured. **Real or constant price variables** adjust nominal variables for changes in the general level of prices. They are inflation-adjusted measures.
 - **Scatter diagrams** show the relationship between two variables plotted in the diagram. By fitting a line through these points we summarize the average relationship between the two variables.
- Econometrics** is the use of statistics by economists to measure relationships between many variables simultaneously. In principle this allows us to get round the '**other things equal**' problem, which always applies in two dimensions.
- Analytical diagrams are often useful in building a model. They show relationships between two variables holding other things equal. If we wish to change one of these other things, we have to shift the line or curve we have shown in our diagram.
 - To understand how the economy works we need both theory and facts. We need theory to know what facts to look for: there are too many facts for the facts alone to tell us the correct answer. Facts without theory are useless, but theory without facts is unsupported assertion. We need both.

Review questions



EASY

- 1 The police research department wants to study whether the level of crime is affected by the unemployment rate. (a) How would you test this idea ? What data would you want ? (b) What 'other things equal' problems would you bear in mind ?
- 2 Suppose the relationship between income and consumption is such that higher incomes are associated with higher consumption of a similar amount. How can the relationship be represented on a graph ?
- 3 Which of the following graphs shows a negative linear relationship ?



4 Suppose the relationship between the price of a car and its quantity demanded, other things equal, is graphed as a downward-sloping straight line. Suppose the income of consumers increases and the demand for cars increases. What happens to the downward-sloping straight line when demand for cars increases owing to an increase in consumers' income?

5 **Common fallacies** Why are these statements wrong? (a) The purpose of a theory is to let you ignore the facts. (b) Economics cannot be a science since it cannot conduct controlled laboratory experiments. (c) People have feelings and act haphazardly. It is misguided to reduce their actions to scientific laws.

MEDIUM

6 Suppose you earn £1200. You purchase only shirts, which are £10 each, with your income every month. (a) What is the purchasing power of your income? (b) Now, suppose your income increases to £1500 a month but the price of shirts remains the same. Does the purchasing power of your income change? (c) Suppose your income remains £1200 but the price of shirts increases to £12 each. Does the purchasing power of your income change?

7 The table below shows car prices and the retail price index (RPI) for the years 2000, 2011, 2012. Calculate the real prices of cars by expressing them at the 2012 prices.

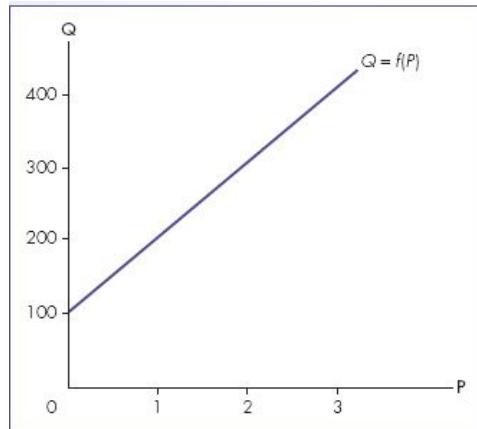
	2000	2012	2012
Car prices (£000s)	6.2	10.4	9.6
RPI (2012 = 100)	9.1	15.4	100

8 The table below shows car prices and house prices for the years 2011 and 2012. Calculate the absolute change and the percentage change in the prices of cars and houses between the years 2011 and 2012.

	2011	2012
Car prices (£)	10 000	9600

	2011	2012
House prices (£)	£200 000	£192 000

9 Calculate the slope of the straight line in the graph provided below. Also, write down the equation of the straight line.



- |0 Why do other influences, particularly the changing level of interest rates, have only a small effect on household decisions about how much income to spend or save ?
- |1 When we use economic data to test an economic theory, we must choose how high to set the bar in our test. If we say that whenever the data depart at all from the prediction of the theory, we will reject most of our theories, which were only approximations in the first place. Conversely, if we only reject a theory when the data are a long way away from the prediction of the theory, we will hardly ever reject any theory. Which of these two possible mistakes is more dangerous ?

HARD

- |2 The table below shows unemployment rates in the capital (London) and the rest of the country. One-third of the national population lives in the capital. Construct an index of national unemployment, treating 2000 as 100. What weights did you use for the two unemployment rates ? Why ?

Unemployment (%)	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
London	7	6	5	4	6	5	4	4	3	4
Rest of country	10	9	8	8	9	8	8	7	7	8

- |3 Revenues at Tom's Pizza Place have been falling in the past few years. The owner, Tom Booker, wants to change his pricing policies to increase revenue. Considering the factors you think would affect Tom's Pizza Place's revenue, construct a revenue model and explain your model.
- |4 **Essay question** Following the introduction of a congestion charge for driving into Central London, traffic levels initially fell by 20 per cent. Over the next few years, traffic reverted almost to its original level. Does this show that the congestion charge failed to reduce congestion ? Even if it did, might it still be a good idea ?

1 For financial variables, like stock prices, data are available even in real time, minute by minute.

2 The price of new houses in the UK reached a peak in 2007, at around £190 000. In 2008 and 2009 the price of new houses decreased quite substantially as a result of the credit crunch.

3 Here, we use the RPI instead of the CPI because in the UK the RPI used to be the main index used to measure inflation and therefore we have a longer series of data for the RPI than for the CPI.

CHAPTER 3

Demand, supply and the market

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 define the concept of a market
- 2 recognize demand and supply curves
- 3 identify equilibrium price and equilibrium quantity
- 4 understand how price adjustment reconciles demand and supply in a market
- 5 analyse what shifts demand and supply curves
- 6 define free markets and markets with price controls
- 7 understand how markets answer what, how and for whom to produce

Society has to find *some* way to decide what, how and for whom to produce. Modern economies rely heavily on markets and prices to allocate resources between competing uses. The interplay of *demand* (representing the behaviour of buyers) and *supply* (representing the behaviour of sellers) determines the quantity produced of a given good or service and the price at which it is bought and sold.

3.1 The market

Shops and fruit stalls physically bring together the buyer and seller. The stock exchange uses intermediaries (stockbrokers), who transact business on behalf of clients. E-commerce is conducted on the Internet. In

supermarkets, sellers choose the price and let customers choose whether or not to buy. Antique auctions force buyers to bid against each other, with the seller taking a passive role.

Although superficially different, these **markets** perform the same economic function. They determine prices that ensure that the quantity buyers wish to buy equals the quantity sellers wish to sell. Price and quantity cannot be considered separately. In fixing the price of a Bentley at 20 times the price of a Fiat, the market for motor cars ensures that production and sales of Fiats greatly exceed the production and sales of Bentleys. These prices guide society in choosing what, how and for whom to produce.

A **market** is a set of arrangements by which buyers and sellers exchange goods and services.

To understand this process more fully, we need to model a typical market. We concentrate the analysis on markets where each participant is small as a fraction of the number of buyers and sellers. The ingredients are market demand (the behaviour of buyers) and market supply (the behaviour of sellers). We can then study how these interact to see how a market works.

3.2 Demand, supply and equilibrium

Demand is not a particular quantity, such as six bars of chocolate, but rather a full description of the quantity of chocolate buyers would purchase at each and every price that might be charged. Suppose we could ask consumers who want to buy chocolate how much they are willing to pay for some different quantities of chocolate bars. The first column of Table 3.1 shows possible prices of chocolate bars. The second column shows the quantities consumers are willing to buy at these prices. Even if chocolate is free - so the price is zero - only a finite amount is demanded. This is plausible since people may become sick from eating too much chocolate. As the price of chocolate rises, chocolate becomes more and more expensive and therefore the quantity demanded falls, *other things equal*. The fact that, as the price of a good or service increases, the quantity demanded of that good or service decreases (other

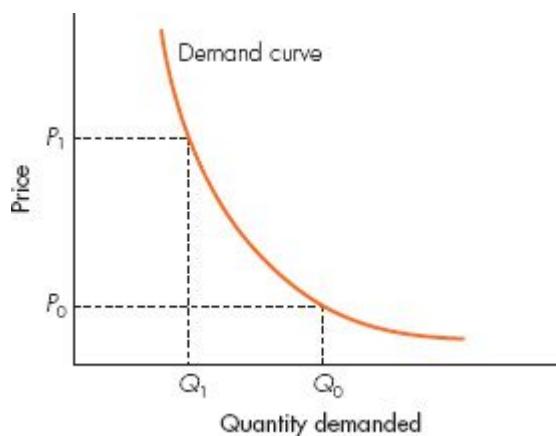
things equal) holds for almost every good or service and is known as the *law of demand*.¹

Demand is the quantity that buyers wish to purchase at each conceivable price.

Table 3.1 Demand and supply of chocolate

(1) Price (£/bar)	(2) Demand (no. of bars)	(3) Supply (no. of bars)
0.00	200	0
0.10	160	0
0.20	120	40
0.30	80	80
0.40	40	120
0.50	0	160

A typical demand curve for a given good is displayed in Figure 3.1. On the vertical axis we plot the price of a good and on the horizontal axis the quantity demanded of that good. As the price of the good increases, say from P_0 to P_1 , the quantity demanded of that good decreases from Q_0 to Q_1 . Notice that in Figure 3.1 the demand curve is not a straight line. In the rest of the book we will consider mainly linear demand curves. Linear demand curves are simpler to analyse than non-linear ones.



A typical demand curve implies a negative relationship between the price of a given good, or service, and the quantity demanded of that good, or service, other things equal.

Figure 3.1 A typical demand curv

Let's consider Table 3.1 in more detail. Suppose that nobody buys any chocolate if the price exceeds £0.40. Together, columns (1) and (2) describe the demand for chocolate as a function of its price.

Supply is not a particular quantity but a complete description of the quantity that sellers want to sell at each possible price. Suppose that we can ask all suppliers of chocolate in the market how much they wish to sell at different prices. The third column of Table 3.1 summarizes the possible behaviour of the sellers. Nobody would supply chocolate at a zero price. In our example, it takes a price of £0.20 before there is an incentive to supply chocolate. At higher prices it is more lucrative to supply chocolate bars and there is a rise in the quantity supplied. This positive relationship between the quantity supplied of a given good or service and the price of that good or service (other things equal) is a regularity that holds for almost every good or service. We call this positive relationship between the price and the quantity supplied of a good or service, the *law of supply*. Together, columns (1) and (3) describe the supply of chocolate bars as a function of their price.

Supply is the quantity of a good that sellers wish to sell at each possible price.



A typical supply curve implies a positive relationship between the price of a given good, or service, and the quantity supplied of that good, or service, other things equal.

Figure 3.2 A typical supply curve

An example of a typical supply curve is given in Figure 3.2. As the price of a good increases, say from P_0 to P_1 , the quantity supplied of that good increases from Q_0 to Q_1 . We plot a supply curve that is not a straight line; however, in the rest of the book we will deal mainly with linear supply curves.

Note the distinction between *demand* and the *quantity demanded*. Demand describes the behaviour of buyers at every price. At a particular price there is a particular quantity demanded. The term ‘quantity demanded’ makes sense only in relation to a particular price. The same applies to *supply* and *quantity supplied*.

In everyday language, we say that when the demand for football tickets exceeds their supply some people do not get into the ground. Economists must be more precise. At the price charged for tickets, the quantity demanded exceeded the quantity supplied. Now suppose that the ticket price increases. The quantity demanded will be reduced. Here there is no change in demand, the schedule describing how many people want admission at each possible ticket price. What has changed is the quantity demanded because the price changed.

The demand and supply schedules are each constructed as a relationship between the quantity demanded and supplied and the price only, keeping ‘other things equal’. In the demand for football tickets, one of the ‘other things’ is whether the game is televised. If it is, the quantity of tickets demanded at each possible price can be lower than if the game is not televised.

Think again about the market for chocolate in Table 3.1. Other things equal, the lower the price of chocolate, the higher the quantity demanded. Other things equal, the higher the price of chocolate, the higher the quantity supplied. A campaign by dentists warning of the effect of chocolate on tooth decay, or a fall in household incomes, would change the ‘other things’ relevant to the demand for chocolate. Either of these changes would reduce the demand for chocolate, reducing the quantities demanded at each price (i.e. the demand curve shifts to the left). Cheaper cocoa beans, or technical advances in packaging chocolate bars, would change the ‘other things’ relevant to the supply of chocolate bars. They would tend to increase the supply of chocolate bars, increasing the quantity supplied at each possible price (i.e. the supply curve shifts to the right).

The equilibrium price

Assume, initially, that all these other things remain constant. We combine the behaviour of buyers and sellers to model the market for chocolate bars. At low prices, the quantity demanded exceeds the quantity supplied but the reverse is true at high prices. At some intermediate price, which we call the **equilibrium price**, the quantity demanded just equals the quantity supplied.

The **equilibrium price** is the price at which the quantity supplied equals the quantity demanded.

Excess demand exists when the quantity demanded exceeds the quantity supplied at the ruling price.

In Table 3.1 the equilibrium price is £0.30, at which 80 bars is the *equilibrium quantity*, the quantity buyers wish to buy and sellers wish to sell. At prices below £0.30, the quantity demanded exceeds the quantity supplied and some buyers are frustrated. There is a shortage. When economists say there is **excess demand** they are using a shorthand for the more accurate expression: the quantity demanded exceeds the quantity supplied *at this price*.

Conversely, at any price above £0.30, the quantity supplied exceeds the quantity demanded. Sellers have unsold stock. Economists describe this surplus as **excess supply**, shorthand for an excess quantity supplied *at this price*. Only at £0.30, the equilibrium price, are quantity demanded and quantity supplied equal. The market clears. People's wishes are fulfilled at the equilibrium price.

Excess supply exists when the quantity supplied exceeds the quantity demanded at the ruling price.

Is a market automatically in equilibrium? What could bring this about? Suppose the price is initially £0.50, above the equilibrium price. Suppliers offer 160 bars but nobody buys at this price. Sellers cut the price to clear their stock. Cutting the price to £0.40 has two effects. It raises the quantity demanded to 40 bars and cuts the quantity producers wish to make and sell to 120 bars. Both effects reduce excess supply.

Price-cutting continues until the equilibrium price of £0.30 is reached and excess supply disappears. At this price, the market clears.

If the price is below the equilibrium price, the process works in reverse. At a price of £0.20, 120 bars are demanded but only 40 are supplied. Sellers run out of stock and charge higher prices. This incentive to raise prices continues until the equilibrium price is reached, excess demand is eliminated and the market clears.

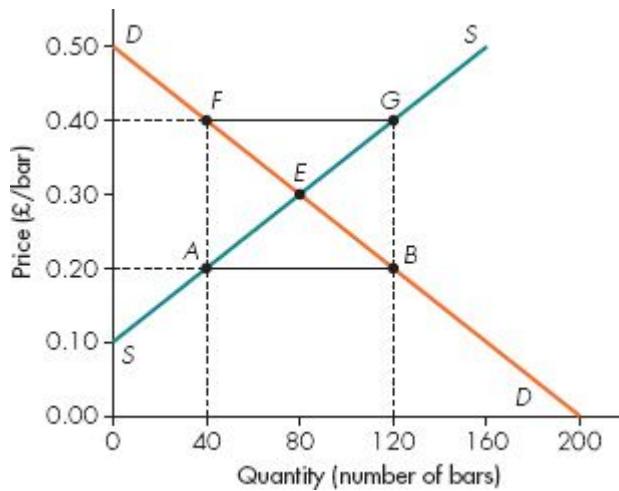
At a particular time, the price may not be the equilibrium price. If not, there is either excess supply or excess demand, depending on whether the price lies above or below the equilibrium price. But these imbalances provide the incentive to change prices towards the equilibrium price. Markets are self-correcting. Some of the key issues in economics turn on how quickly prices adjust to restore equilibrium in particular markets.

3.3 Demand and supply curves

Table 3.1 shows demand and supply conditions in the chocolate market and allows us to find the equilibrium price and quantity. It is useful to analyse the same problem diagrammatically.

Figure 3.3 measures chocolate prices on the vertical axis and chocolate quantities on the horizontal axis. The **demand curve** *DD* plots the data in the first two columns of Table 3.1 and joins up the points. In this case the demand curve is a straight line, though it need not be. Although the demand is a straight line we continue to label it as a demand curve.

Our demand has a negative slope. Larger quantities are demanded at lower prices.



Market equilibrium is at E. At prices below the equilibrium price there is excess demand: AB shows the excess demand at the price £0.20. At prices above the equilibrium price there is excess supply: FG shows the excess supply at the price £0.40.

Figure 3.3 The market for chocolate

Figure 3.3 plots columns (1) and (3) of Table 3.1. Joining up the different points yields the **supply curve** SS. Again, this happens to be a straight line but it need not be. It slopes up because suppliers only wish to increase the quantity supplied if they get a higher price.

The **demand curve** shows the relationship between price and quantity demanded, other things equal.

The **supply curve** shows the relationship between price and quantity supplied, other things equal.

We can now re-examine excess supply, excess demand and equilibrium. A particular price is shown by a height on the vertical axis, a particular quantity by a length on the horizontal axis. Equilibrium is at point E. As in Table 3.1, this entails a price of £0.30 and a quantity of 80 bars. At any price below the equilibrium price, the horizontal distance between the supply curve and the demand curve is the excess demand at that price. At £0.20, 40 bars are supplied but 120 bars are demanded. The distance AB is the excess demand of 80 bars. Conversely, above the equilibrium price there is excess supply. At £0.40, 40 bars are demanded,

120 bars are supplied and the horizontal distance FG is the excess supply of 80 bars at this price.

Suppose the price is £0.40. Only 40 bars are sold, even though sellers would like to sell 120 bars. Why are sellers - not buyers - frustrated when their wishes differ? Participation in a market is voluntary. Buyers are not *forced* to buy nor sellers *forced* to sell. When markets are not in equilibrium, the quantity transacted is the *smaller* of the quantity supplied and the quantity demanded. Any quantity above 40 bars at a price of £0.40 would force buyers into purchases they do not want. Similarly, at a price of £0.20, any quantity greater than 40 bars involves sellers in forced sales.

We can now reconsider *price determination* in the chocolate market. Figure 3.3 implies that there is excess supply at all prices above the equilibrium price of £0.30. Sellers react to unsold stocks by cutting prices. The existence of an excess of supply creates a downward pressure on the price; therefore the price falls. Once the price falls to the equilibrium price, excess supply is eliminated. Equilibrium is at point E . Conversely, at prices below £0.30 there is excess demand. The existence of an excess of demand creates an upward pressure on the price; therefore the price increases. The increase in the price eliminates the excess of demand until the equilibrium point E is reached. In equilibrium, buyers and sellers can trade as they wish at the equilibrium price. There is no incentive for any further price changes.

MATHS 3.1

MARKET EQUILIBRIUM WITH LINEAR DEMAND AND SUPPLY

We can describe the equilibrium in a given market in a simple mathematical way. First, we introduce the *direct demand function* as a relationship between the quantity demanded and the price of a given good or service, keeping ‘other things constant’. Suppose that the direct demand function is linear; it can be written as

$$Q^D = a - bP \quad (1)$$

where Q^D denotes the quantity demanded, P the price and a and b are two positive constants. Using letters instead of numbers makes

the analysis a bit more general. Equation (1) implies a negative relationship between the quantity demanded and the price of a given good or service.

Next, we introduce a linear *direct supply function*:

$$Q^S = c + dP \quad (2)$$

where Q^S is the quantity supplied and c and d are two constants. We assume that the constant d is positive. This implies that there is a positive relationship between the quantity supplied and the price. The constant c can be positive or negative.

The market equilibrium is where quantity demanded equals quantity supplied, meaning

$$Q^D = Q^S$$

This fact implies that

$$a - bP = c + dP \quad (3)$$

Solving equation (3) for the price P gives us:

$$\begin{aligned} dP + bP &= a - c \\ \Rightarrow P(b + d) &= a - c \\ \Rightarrow P^* &= \frac{(a - c)}{(b + d)} \end{aligned}$$

P^* is the equilibrium price that equates quantity demanded and quantity supplied.

To find the corresponding equilibrium quantity we can substitute the expression of P^* into the original demand function or into the supply function. Here we use the demand function:

$$\begin{aligned}
 Q^D &= a - b \frac{(a - c)}{(b + d)} \\
 \Rightarrow Q^D &= \frac{a(b + d) - b(a - c)}{(b + d)} \\
 \Rightarrow Q^D &= \frac{ba + da - ba + bc}{(b + d)} \\
 \Rightarrow Q^* &= \frac{bc + da}{(b + d)}
 \end{aligned}$$

Therefore the market equilibrium in our example is given by:

$$P = \frac{(a - c)}{(b + d)} \quad \text{and} \quad Q^* = \frac{bc + da}{(b + d)}$$

From the direct demand function we can always find the *inverse demand function* (and vice versa). The inverse demand function is a relationship between the price and the quantity demanded of a given good. The inverse demand function associated with equation (1) is given by

$$P = \frac{a}{b} - \frac{1}{b} Q^D \tag{4}$$

Why do we need an inverse demand? Because when we plot a demand function on a graph we put the price on the vertical axis and the quantity on the horizontal axis. Therefore, we normally plot the inverse demand. This is what we have called the demand curve. The same applies for the supply function. The *inverse supply function* associated with equation (2) is given by

$$P^* = -\frac{c}{d} + \frac{1}{d} Q^S \tag{5}$$

3.4 Behind the demand curve

The demand curve depicts the relationship between price and quantity demanded *holding other things constant*. What are those ‘other things’? The other things relevant to demand curves can usually be grouped under three headings: the price of related goods, the income of consumers (buyers) and consumer tastes or preferences. We look at each of these in turn.

The price of related goods

In Chapter 2 we discussed the demand for tube travel. A rise in bus fares or petrol prices would increase the quantity of tube travel demanded at each possible tube price. In everyday language, buses and cars are *substitutes* for the tube. Similarly, petrol and cars are *complements* because you cannot use a car without also using fuel. A rise in the price of petrol tends to reduce the demand for cars.

How do substitutes and complements relate to the demand for chocolate bars? Clearly, other sweets (jelly babies, say) are *substitutes* for chocolate. An increase in the price of other sweets increases the quantity of chocolate demanded at each possible chocolate price, as people substitute away from other sweets towards chocolate. If people buy chocolate to eat at the cinema, films would be a *complement* to bars of chocolate. A rise in the price of cinema tickets would reduce the demand for chocolate since fewer people would go to the cinema. Nevertheless, it is difficult to think of many goods that are complements to chocolate. Complementarity is indeed a more specific feature than substitutability (DVD players and DVDs, coffee and milk, shoes and shoelaces).

A price increase for one good raises the demand for *substitutes* for this good but reduces the demand for *complements* to the good.

Consumer incomes

The second category of ‘other things equal’ when we draw a particular demand curve is consumer income. When incomes rise, the demand for most goods increases. Typically, consumers buy more of everything. However, there are exceptions.

As their name suggests, most goods are *normal goods*. *Inferior goods* are typically cheap but low-quality goods that people prefer not to buy if they can afford to spend a little more.

For a *normal good*, demand increases when incomes rise. For an *inferior good*, demand falls when incomes rise.

Tastes and expectations

The third category of things held constant along a particular demand curve is consumer tastes and their expectations. In part, tastes are shaped by convenience, custom and social attitudes. The fashion for the mini-skirt reduced the demand for fabric. The emphasis on health and fitness has increased the demand for jogging equipment, health foods and sports facilities while reducing the demand for cream cakes, butter and cigarettes. Also, what consumers expect about future prices affects their demand. If consumers expect the price of a good to increase in the future they will start buying it now, and so demand for that good will increase.

CASE 3.1

HORSEMEAT BURGER? NO, THANKS!

In February 2012 it was reported in the media that, in many European countries, horse DNA had been found in various frozen processed meat items, such as beefburgers and ready meals.

While there is no health risk involved in eating horsemeat, we would expect consumers to reduce their demand for those frozen products. On the one hand, this reduction in demand has to do with taste. Many consumers did not like the idea of eating horsemeat. On the other hand, the horsemeat scandal intensified concerns about the quality of cheap food. Consumers now felt that the quality of frozen burgers and ready meals was too uncertain.

We can use our simple demand and supply model to see what should happen to the market of frozen processed meat in this case. The horsemeat scandal represents a change in consumers' tastes and expectations. Demand should shift to the left while supply should remain unchanged. This implies that both the quantity and the price of frozen processed meat should decrease. On the other hand, we should expect demand for related goods, like freshly prepared burgers made by local butchers, to increase. Has this happened? Let's look at the UK market.

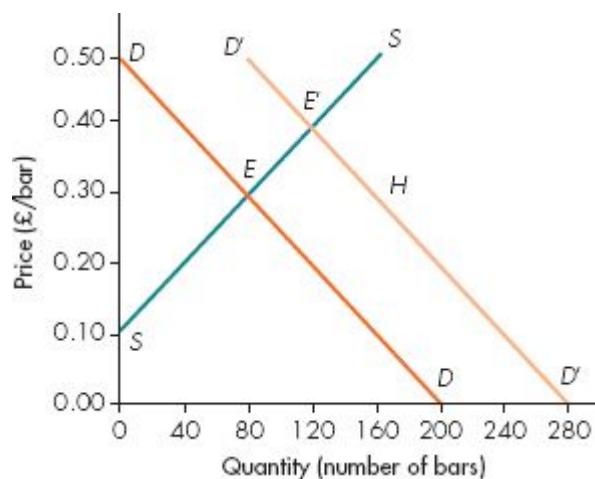
According to the *Financial Times*, sales of frozen burgers plunged by 43 per cent in the four weeks after the horsemeat scandal became public. Sales of frozen ready meals were down by 13 per cent in the same period. On the other hand, the demand for

processed meat prepared by local butchers increased by 30 per cent in the same period according to the BBC.

Sources: adapted from ‘Sales of frozen burgers plunge’, *Financial Times*, 26 February 2013; ‘Horsemeat scandal: Good news for butchers?’, BBC News, 13 February 2013.

3.5 Shifts in the demand curve

We can now distinguish between movements along a given demand curve and shifts in the demand curve itself. In Figure 3.3 we drew the demand curve for chocolate bars for a given level of the three underlying factors: the price of related goods, incomes and tastes. Movements along the demand curve isolate the effects of chocolate prices on quantity demanded, holding other things equal. Changes in any of these three factors will change the demand for chocolate. In particular, any change in those three factors will *shift* the demand for chocolate.



At low ice cream prices, the demand curve for chocolate is DD and the market equilibrium occurs at the point E. Higher ice cream prices raise the demand for chocolate, shifting the demand curve to D'D'. At the former equilibrium price there is now excess demand EH, which gradually bids up the price of chocolate until the new equilibrium is reached in E'.

Figure 3.4 An increase in chocolate demand

Figure 3.4 shows a rise in the price of a substitute for chocolate, say ice cream, which leads people to demand more chocolate and less ice cream. When the price of ice cream increases at each chocolate price, there is

now a larger quantity of chocolate demanded. People substitute chocolate for ice cream. This *shifts* the demand curve for chocolate from DD to $D'D'$. The entire demand curve shifts to the right. At each price on the vertical axis, a larger horizontal distance indicates a higher quantity demanded.

Changes in the price of ice cream have no effect on the incentives to supply chocolate bars: at each price of chocolate, suppliers wish to supply the same quantity of chocolate as before. The increase in demand, or rightward shift in the demand curve, changes the equilibrium price and quantity in the chocolate market. Equilibrium has changed from E to E' . The new equilibrium price is £0.40 and the new equilibrium quantity is 120 bars.

We can sketch the transition from the old equilibrium at E to the new equilibrium at E' . When ice cream prices first rise, the demand curve for chocolate shifts from DD to $D'D'$. With the chocolate price still at £0.30, there is an excess demand EH : 160 bars are demanded but only 80 bars are supplied. This excess demand bids up prices, which gradually rise to the new equilibrium price of £0.40, choking the quantity demanded back from 160 bars to 120 bars and providing the incentive to raise the quantity supplied from 80 bars to 120 bars.

We draw two lessons from this example. First, the quantity demanded depends on four things: its own price, prices of related goods, incomes and tastes. We could draw a two-dimensional diagram showing the relation between quantity of chocolate demanded and any one of these four things. The other three things would then be the ‘other things equal’ for this diagram. In drawing demand curves, we single out the price of the commodity itself (here, the price of chocolate bars) to put in the diagram with quantity demanded. The other three factors are the ‘other things equal’ for drawing a particular demand curve. Changes in any of these other three things shift the position of demand curves.

Why single out the price of the commodity itself to plot against quantity demanded? We want to study the market for chocolate. Prices of related goods, incomes and tastes are determined elsewhere in the economy. By focusing on the price of chocolate, we see the self-correcting mechanism by which the market reacts to excess demand or excess supply: inducing changes in chocolate prices within the chocolate market restores equilibrium.

Second, our example illustrates analysis by *comparative statics*. The analysis is comparative because it compares the old and new equilibria, and static because it compares only the equilibrium positions. In each equilibrium, prices and quantities are constant. **Comparative static analysis** is not interested in the dynamic path by which the economy moves from one equilibrium to the other; it is interested only in the point from which it began and the point at which it ends.

Comparative static analysis changes one of the ‘other things equal’ and examines the effect on equilibrium price and quantity.

Using Figure 3.4 we can also analyse a change in one of the ‘other things equal’. Suppose the demand curve is initially $D'D'$ and the market begins in equilibrium at E' . Then the demand for chocolate falls to DD . This might reflect a fall in the price of a chocolate substitute, a fall in consumer incomes or a change in tastes away from liking chocolate. When the demand curve shifts left to DD , showing less chocolate demanded at each price, the new equilibrium is at E . At the original price of £0.40 there is excess supply, which bids prices down to the new equilibrium price of £0.30. When the demand curve shifts to the left, there is a fall in both equilibrium price and equilibrium quantity.

CASE 3.2

THE DEMAND FOR MOBILE APPS

We have defined demand as quantity that buyers are willing to buy at each conceivable price. To construct a market demand for a given good using our definition, we should be able to ask each consumer in the market how much she wishes to buy of the good at each possible price. As you can guess, it is very impractical to obtain such information. Instead, we can use other information that is easily available to us to obtain market demand data. We observe the quantity purchased of a good in a given period of time and the price of the good when the purchase was made. We can observe the prices of other goods that we think are substitutes for or complements to the good we are analysing, and so on. We can use those data to *estimate* the market demand for a given good.

Estimating economic relationships using economic data is the main objective of an important branch of economics known as *econometrics*. We are not interested in explaining how econometrics works, we just want to outline that, from available economic data, it is possible to obtain a numerical expression for a market demand for any particular good we may be interested in. Here we present an example of an estimated demand for mobile apps in 2011. Data come from two leading app stores, Apple App Store and Google Android Market. The data are from the South Korean market and the sample contains 23 145 observations. The estimated demand is:

$$Q = 22.334 + 0.282P + 0.339AppSize + 1.224AppAge - 1.425AppLength \quad (1)$$

where Q is the quantity of mobile apps downloaded during the period, P is the price of the apps, $AppSize$ is the size of the application in mega bytes, $AppAge$ is the time elapsed since the application was first launched and $AppLength$ is the length of textual app description. The demand equation includes the quantity demanded, the price and other variables that may be relevant in affecting the demand for applications.

From equation (1) we can see that an increase in the price of apps (other things equal) decreases the quantity demanded of applications, as we would expect from market demand. Demand increases with the application size (other things equal). Longer is the textual description of the applications (other things equal) and lower is the demand.

The estimated demand in (1) represents useful information for companies creating mobile applications. For example, it can be used to forecast what the demand of a mobile application is going to be if its price increases.

A word of caution about the estimated demand in (1): econometrics uses statistical techniques to estimate a relationship between variables using data. Therefore equation (1) is not the exact market demand and the predictions we obtain from it may not be totally reliable.

Source: adapted from A. Ghose and S. P. Han, 'Estimating demand for mobile apps in the new economy', proceedings of the International Conference on Information Systems

3.6 Behind the supply curve

At low prices, only the most efficient chocolate producers make profits. As prices rise, producers previously unable to compete can now make a profit in the chocolate business and wish to supply. Moreover, previously existing firms may be able to expand output by working overtime, or buying fancy equipment unjustified when selling chocolate at lower prices. In general, higher prices are needed to induce firms to produce more chocolate. Other things equal, supply curves slope up as we move to the right.

Just as we studied the ‘other things equal’ along a demand curve, we now examine four categories of ‘other things equal’ along a supply curve: the technology available to producers, the cost of inputs (labour, machines, fuel, raw materials), government regulation and expectations. Along a particular supply curve, all of these are held constant. A change in any of these categories shifts the supply curve, changing the amount producers wish to supply at each price.

Technology

A supply curve is drawn for a given technology. Better technology shifts the supply curve to the right. Producers supply more than previously at each price. Better cocoa refining reduces the cost of making chocolate. Faster shipping and better refrigeration lead to less wastage in spoiled cocoa beans. Technological advance enables firms to supply more at each price.

As a determinant of supply, technology must be interpreted broadly. It embraces all know-how about production methods, not merely the state of available machinery. In agriculture, the development of disease-resistant seeds is a technological advance. Improved weather forecasting might enable better timing of planting and harvesting. A technological advance is any idea that allows more output from the same inputs as before. In the terminology of Chapter 1, a technological advance shifts the production possibility frontier outwards.

Input costs

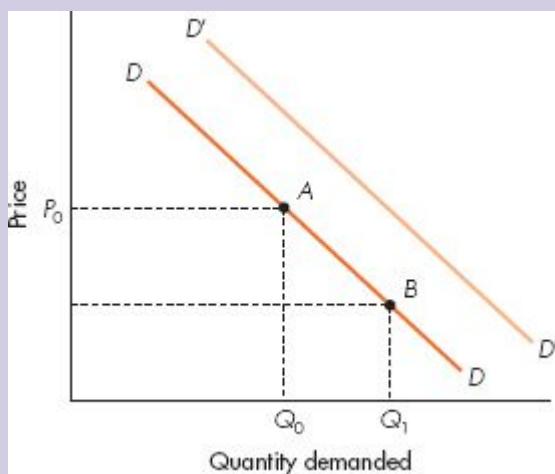
A particular supply curve is drawn for a given level of input prices. Lower input prices (lower wages, lower fuel costs) induce firms to supply more output at each price, shifting the supply curve to the right. Higher input prices make production less attractive and shift the supply curve to the left. If a late frost destroys much of the cocoa crop, scarcity will bid up the price of cocoa beans. Chocolate producers supply less chocolate at each price than previously.

ACTIVITY 3.1

MOVEMENT ALONG A CURVE VS SHIFTS OF THE CURVE

From the initial point A , the figure below shows two quite different ‘increases in demand’. One is an increase in the quantity demanded, from Q_0 to Q_1 , moving along the curve from A to B . This is the effect of a price cut but *not* an increase in demand since the demand curve DD is unaffected.

By an increase in demand, we mean a shift in the demand curve, say from DD to $D'D'$, which also increases quantity demanded from Q_0 to Q_1 at the going price P_0 . This shift in demand reflects an increase in the price of a substitute good (decrease in the price of a complementary good), an increase in income (assuming it is a normal good) or a change in taste.



Similarly, sellers adjust to higher prices by moving up a given supply curve. But an increase in supply means a shift to the right of

the whole supply curve, caused by lower input prices, new technology or less regulation.

Other things equal, changes in price move us *along* demand and supply curves. When other determinants change, they shift these schedules.

Questions

Classify each of the following as an upward or downward shift in the supply or demand curve:

- a. New interactions between Europe and China make wages of unskilled European workers fall.
- b. New interactions between Europe and China make the price of European coal increase.
- c. The government ban on city parking by large cars reduces the price of Bentleys.

To check your answers to these questions, go to page 677.

Government regulation

In discussing technology, we spoke only of technological advances. Once people have discovered a better production method they are unlikely subsequently to forget it.

Government regulations can sometimes be viewed as imposing a technological change that is *adverse* for producers. If so, the effect of regulations will be to shift the supply curve to the left, reducing quantity supplied at each price.

More stringent safety regulations prevent chocolate producers using the most productive process because it is quite dangerous for workers. Anti-pollution devices may raise the cost of making cars, and regulations to protect the environment may make it unprofitable for firms to extract surface mineral deposits which could have been cheaply quarried but whose extraction now requires expensive landscaping. Whenever regulations prevent producers from selecting the production methods they would otherwise have chosen, the effect of regulations is to shift the supply curve to the left.

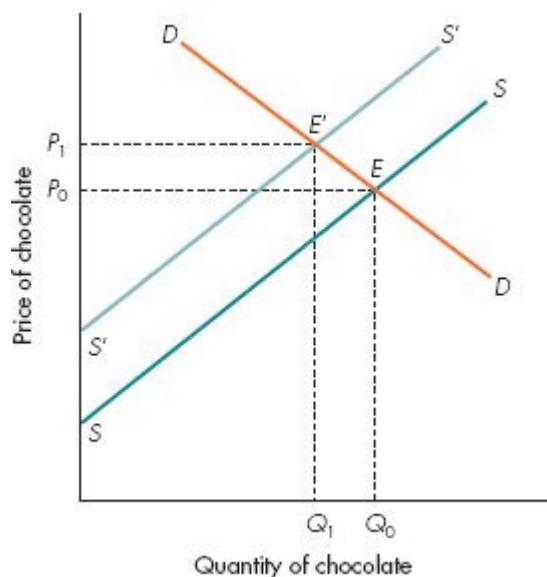
Expectations

If a firm expects the price of its product to fall in the future it has an incentive to supply more today. Firms' expectations about future price changes affect their current supply decisions.

3.7 Shifts in the supply curve

Along a given supply curve, we hold constant technology, the prices of inputs and the extent of government regulation. Any change in those factors will *shift* the supply curve. We now undertake a comparative static analysis of what happens when a change in one of these 'other things equal' leads to a fall in supply. Suppose tougher safety legislation makes it more expensive to make chocolate bars in mechanized factories. Figure 3.5 shows a shift to the left in the supply curve, from SS to $S'S'$. Equilibrium shifts from E to E' .

The equilibrium price *rises* but equilibrium quantity *falls* when the supply curve shifts to the left. Conversely, a rise in supply shifts the supply curve from $S'S'$ to SS . Equilibrium shifts from E' to E . A rise in supply induces a *higher* equilibrium quantity and *lower* equilibrium price.



The supply curve initially is SS and market equilibrium is at E . A reduction in the supply of chocolate shifts the supply curve to the left to $S'S'$. The new equilibrium at E' has a

higher equilibrium price and a lower equilibrium quantity than the old equilibrium at E .

Figure 3.5 A fall in supply

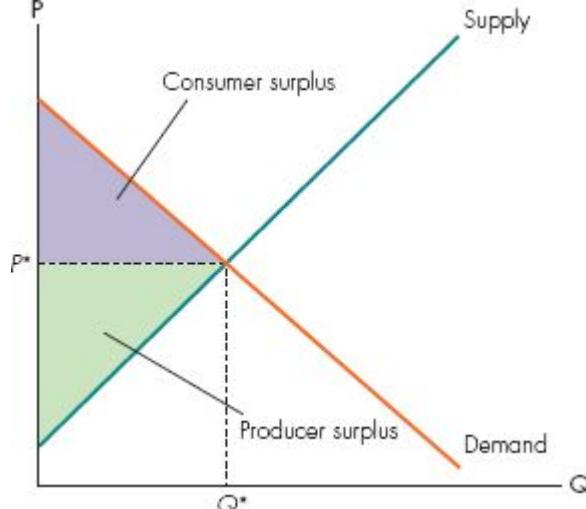
3.8

Consumer and producer surplus

In previous sections we defined the market equilibrium. Can we say something about how ‘good’ a market equilibrium is? In practice, we want to find a possible measure for the gains that consumers and sellers obtain from trading at the equilibrium price. For the consumers, this measure of trade gain is called *consumer surplus*.

For a single consumer, the consumer surplus is the difference between the maximum price (also called the *reservation price*) that she is willing to pay for a given amount of a good or service and the price she actually pays.

Suppose that you want to buy the box set of your favourite TV series. You are willing to pay a maximum £30 for it. If the price of the box set in a shop is £15, you buy it, and you can say that from buying it you have obtained a surplus of £15. This surplus is a measure of your gain from buying the box set.



The total gain from trading in the market is given by the sum of the consumer and producer surplus. Here, the gain from trading at the equilibrium price is depicted.

Figure 3.6 Consumer and producer surplus at the market equilibrium

We can extend the idea of consumer surplus to all consumers in a market. In particular, the consumer surplus is measured by the area below the market demand curve and above the equilibrium price.

Similarly we can define a measure for the gain sellers obtain from selling a given quantity of a good or service at the equilibrium price. We call this gain for sellers the *producer surplus*. The producer surplus for sellers is the amount that sellers benefit by selling at a market price that is higher than they would be willing to sell for.

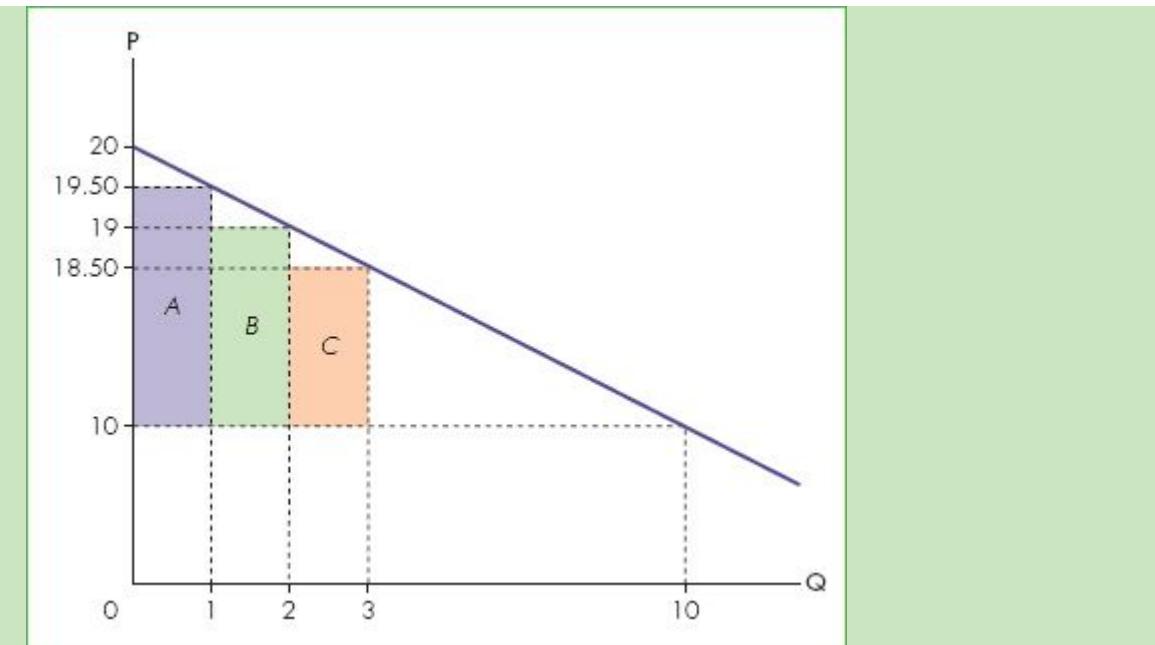
Suppose you want to sell an old vinyl record on eBay. You are willing to sell it at a minimum price of £10. Suppose you end up selling it for £30. Your surplus from this transaction is £20. Graphically, the producer surplus is given by the area above the market supply and below the equilibrium price.

The sum of the consumer and producer surplus in a market is a measure of the economic surplus that the participants obtain by trading in the market. This is shown in Figure 3.6. It should be noticed that the economic surplus is highest at the equilibrium price. At any price that is not the equilibrium price, the economic surplus will be lower.

CONCEPT 3.1

GRAPHICAL DERIVATION OF CONSUMER AND PRODUCER SURPLUS

Consider a linear market demand for a given good. Suppose that the equilibrium price that consumers pay is £10 for each unit of the good and the equilibrium quantity is 10 units. Suppose that consumers are willing to buy one unit of the good at a price of £19.50. They are willing to buy two units of the good if the price of each unit is £19. They are willing to buy three units if the unit price is £18.50, and so on. In the figure on the right we plot market demand with the information just described.



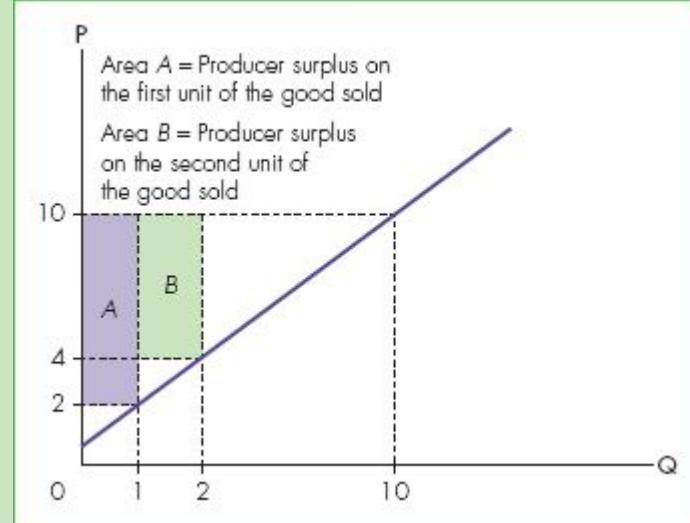
Area A = Consumer surplus on the first unit of the good

Area B = Consumer surplus on the second unit of the good

Area C = Consumer surplus on the third unit of the good

Consumers are willing to buy 1 unit of the good at the price of £19.50; however, they actually pay £10 for each unit of the good.

Therefore, their gain from trading at the equilibrium price is £9.50 on the first unit of the good. This is given by area *A* in the figure above. The consumers are willing to buy two units of the good if the price is £19 for each unit. Since they pay £10 for each unit, their gain is £9 on the second unit of the good. This is given by area *B* in the figure above. Therefore, the surplus obtained by the consumers from buying the first two units of the good is given by the sum of area *A* and area *B*. We can continue this process until we arrive at the equilibrium quantity 10. The total consumer surplus will be approximately given by the area below the market demand and above the equilibrium price. We have an approximation because we have considered a good that can be traded only in discrete units (1, 2, 3, and so on). If we assume that the good can be sold in any possible amount (not only discrete), then the consumer surplus is exactly the entire area below the market demand and above the equilibrium price.



Area A = Producer surplus on the first unit of the good sold

Area B = Producer surplus on the second unit of the good sold

For the producer surplus, we can use similar reasoning. Suppose that the equilibrium price is £10 and the quantity sold in equilibrium is ten. At a price of £2, the sellers are willing to supply one unit of the good. At a price of £4, they are willing to supply two units of the good. In the figure on the right we show the market supply for this case.

By selling ten units of the good, on the first unit the sellers obtain a surplus of £8 (£10 2 £2). This is represented by area *A* in the figure. On the second unit sold, the surplus obtained by the sellers is £6 (£10 2 £4). This is represented by area *B*. Therefore the surplus obtained by selling two units of the good is given by area *A* plus area *B*. We can continue this process until we reach the equilibrium quantity. The producer surplus is approximately given by the area above the market supply and below the equilibrium price. Again, if we allow the good to be sold in any possible amount the approximation will be exact.

3.9 Free markets and price controls

Government actions may shift demand and supply curves, as when changes in safety legislation shift the supply curve, but in a **free market**

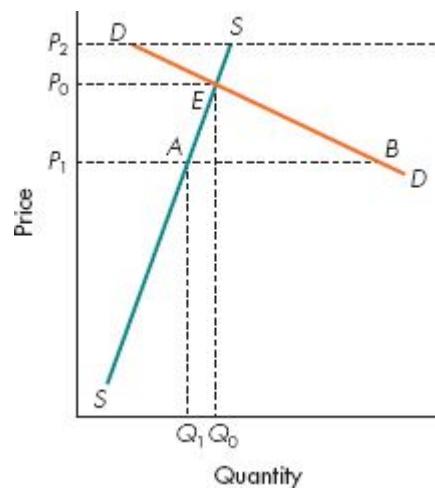
the government makes no attempt to regulate prices directly. If prices are sufficiently flexible, the pressure of excess supply or excess demand will quickly bid prices in a free market to their equilibrium level. Markets will not be free when effective **price controls** exist. When price controls are in place in a market, the economic surplus of the participants in that market will also change. Price controls may be *floor* prices (minimum prices) or *ceiling* prices (maximum prices).

Free markets allow prices to be determined purely by the forces of supply and demand.

Price controls are government rules or laws setting price floors or ceilings that forbid the adjustment of prices to clear markets.

Price ceilings make it illegal for sellers to charge more than a specific maximum price. Ceilings may be introduced when a shortage of a commodity threatens to raise its price a lot (such as food prices during a war). High prices are the way a free market rations goods in scarce supply. This solves the allocation problem, ensuring that only a small quantity of the scarce commodity is demanded, but may be thought unfair, a normative value judgement. High food prices mean hardship for the poor.

Faced with a national food shortage, a government may impose a price ceiling on food so that poor people can afford to eat.



Free market equilibrium occurs at the point *E*. The high price P_0 chokes off quantity demanded to ration scarce supply. A price ceiling at P_1 succeeds in holding down the

price but leads to excess demand AB . It also reduces quantity supplied from Q_0 to Q_1 . A price ceiling at P_2 is irrelevant since the free market equilibrium at E can still be attained.

Figure 3.7 The effect of a price ceiling

Figure 3.7 shows the market for food. Suppose a war has disrupted imports of food. The supply curve is far to the left and the free market equilibrium price P_0 is very high. Instead of allowing free market equilibrium at E , the government imposes a price ceiling P_1 . The quantity sold is then Q_1 and excess demand is the distance AB . The price ceiling creates a shortage of supply relative to demand by holding food prices below their equilibrium level.²

The ceiling price P_1 allows the poor to afford food but it reduces total food supplied from Q_0 to Q_1 . With excess demand AB at the ceiling price, rationing must be used to decide which potential buyers are actually supplied. This rationing system could be arbitrary. Food suppliers may sell supplies to their friends, not necessarily the poor, or may take bribes from the rich who jump the queue.

Holding down the price of food may not help the poor after all. Ceiling prices are often accompanied by government-organized rationing by quota to ensure that available supply is shared out fairly, independent of ability to pay.

CASE 3.3

RENT CEILING IN SWEDEN

The main justification for introducing a rent ceiling is the right to housing.

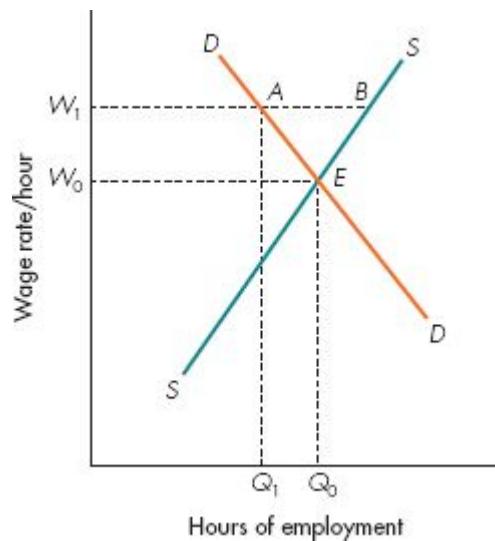
A low rent price is believed to make housing affordable for poor people. Unfortunately, as outlined in the analysis above, the introduction of a rent ceiling may have perverse effects and poor people may not be able to get a cheap house anyway. Sweden provides a very interesting example of a rent control policy. In Sweden, rent price is kept particularly low. A study made by the European University Institute (EUI) showed that:

- (a) To make a 5 per cent return on investment, a Swedish developer would need to set rents 70 per cent higher than allowed by the rent control.
- (b) Rents are little influenced by location, so that metropolitan units are especially underpriced.

The result of this rent control system in Sweden was a reduction in the supply of new properties intended for rental in the market. Of the approximately 30 000 dwellings completed in 2006, only 36 per cent were intended for rental. In comparison, from 1990 to 1996 more than 50 per cent of new dwellings completed were intended for rental. This result is consistent with the analysis we have just made. A price ceiling below the market equilibrium price has the effect of reducing the market supply. This creates a shortage of rental units.

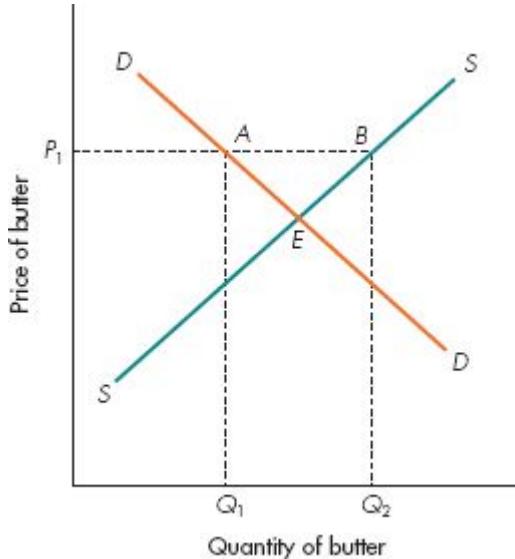
Furthermore, a rent ceiling may have even more perverse effects. Rent control may discourage landlords from maintaining and repairing units during the tenancy. In some cases, landlords collect key money to offset the losses associated with a low rent. This implies that those willing to pay more will get the rental unit, thus eliminating the positive effect of rent control on poor people.

Source: adapted from Prince Christian Cruz, *The pros and cons of rent control* (<http://www.globalpropertyguide.com/investment-analysis/The-pros-and-cons-of-rent-control>).



The demand curve for hours DD and the supply curve of hours SS imply free market equilibrium at E. A legal minimum wage at W_1 raises hourly wages for those who remain employed but reduces the quantity of hours of employment available from Q_0 to Q_1 .

Figure 3.8 A minimum wage



At the floor price P_1 supply is Q_2 , but demand only Q_1 . Only Q_1 will be traded. By buying up the excess supply AB, the government can satisfy both suppliers and consumers at the price P_1 .

Figure 3.9 A price floor for butter

Whereas the aim of a price ceiling is to reduce the price for consumers, the aim of a floor price is to raise the price for suppliers. One example of a floor price is a national minimum wage. In the UK a minimum hourly wage rate was introduced in 1999. Figure 3.8 shows the demand curve and supply curve for labour. The demand for labour tells us for each possible wage rate how many working hours firms demand. The labour supply tells us for each possible wage how many hours workers are willing to work.

The free market equilibrium is at E, where the wage is W_0 . A minimum wage below W_0 is irrelevant since the free market equilibrium can still be attained. Suppose, in an effort to help workers, the government imposes a minimum wage at W_1 . Firms demand a quantity of labour Q_1 and there is excess supply AB. The lucky workers who manage to get work are

better off than before but some workers are worse off since total hours worked fall from Q_0 to Q_1 .

Many countries set floor prices for agricultural products. Figure 3.9 shows a floor price P_1 for butter. In previous examples we assumed that the quantity traded would be the smaller of quantity supplied and quantity demanded at the controlled price, since private individuals cannot be forced to participate in a market. There is another possibility: the government may intervene not only to set the control price but also to buy or sell quantities of the good to supplement private purchases and sales.

CASE 3.4

MORE ON PRICE CONTROLS

California, home of Silicon Valley and Hollywood, is one of the richest places on earth. Yet in 2001 California suffered blackouts as electricity supplies ran out. Since poverty cannot be blamed, it must have been the result of poor policies. California privatized state electricity companies but then capped the price they could charge for electricity. However, the level of the price cap was far too low. Local electricity suppliers haemorrhaged money. This low cap caused the bankruptcy of Pacific Gas and Electric Company (PG&E) and the near bankruptcy of Southern California Edison in early 2001. Not only does an artificially low price lead sooner or later to a lower quantity supplied, it also raises the quantity demanded. Those two effects together were the basis of the electrical blackouts in California. It was estimated that the cost of that electricity crisis was between \$40 billion and \$45 billion.

Another example of a price control policy is the minimum price (a price floor) that the European Commission imposed on Chinese frozen strawberries in 2007.

Why impose such a price floor? Frozen strawberries from China were too cheap compared to the ones produced in Europe. The price floor aimed to punish Chinese exporters for selling the frozen fruit, used in jam and yoghurt, in Europe below domestic prices or below the production cost, a practice known as *dumping*. In this case, the

price floor aimed to protect the European producers of frozen strawberries from Chinese competition.

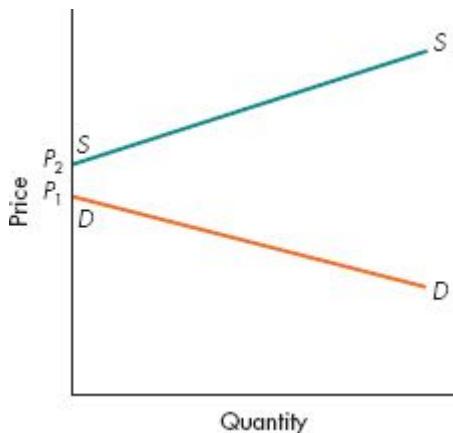
Here, European producers gained from this policy while Chinese exporters probably lost. What about consumers? A price floor, as shown in Figure 3.9, increases the final price paid by the consumer. Therefore consumers were probably worse off as a result of this policy. Nevertheless, the European Commission decided to impose such a price floor, since it believed that the adverse effects on European producers and farmers would be of a substantial and lasting nature should the price floor not be imposed.

3.10 What, how and for whom

The free market is one way for society to solve the basic economic questions of what, how and for whom to produce. In this chapter we have begun to see how the market allocates scarce resources among competing uses.

The market decides how much of a good should be produced by finding the price at which the quantity demanded equals the quantity supplied. Other things being equal, the more of a good is produced in market equilibrium, the higher the quantity demanded at each price (the further the demand curve lies to the right) and the higher the quantity supplied at each price (the further the supply curve lies to the right).

The market tells us for whom the goods are produced: the good is purchased by all those consumers willing to pay at least the equilibrium price for it. The market also tells us who is producing: all those willing to supply at the equilibrium price. Later in this book we shall see that the market also tells us how goods are produced.



Even P_1 , the highest price consumers will pay, is lower than P_2 , the minimum price producers require to produce any of this good.

Figure 3.10 A good not produced

Finally, the market determines what goods are being produced. Nature supplies goods free of charge. People engage in costly production activities only if they are paid. The supply curve tells us how much has to be paid to create supply. Figure 3.10 shows a good that will not be produced. The highest price P_1 that consumers are prepared to pay is still insufficient to persuade producers to produce.

Society may not like the answers the market provides. Free markets *do not* provide enough food to remove hunger or enough medical care to treat all the sick. They provide food and medical care for those willing and *able to pay* the equilibrium price. Society may adopt the normative judgement that the poor should get more food and medical care than they get in a free market. Society may also adopt the normative judgement that, although people are willing and able to pay for pornography, it is socially better to ban some of these activities. Few societies allow unrestricted free markets for all commodities. Governments intervene widely to alter market outcomes, through direct regulation, taxation and transfer payments such as unemployment benefit.

CASE 3.5

ANATOMY OF PRICE AND QUANTITY CHANGES

How should we interpret the figure below showing data for the UK housing industry? What was happening? Was it a shift in demand, in supply or in both that caused this pattern during 1985-2012?

Suppose all the observations represent *equilibrium* prices and quantities in each year. Thus each point reflects the intersection of the demand and supply curve that year. What changes in the ‘other things equal’ determinants of supply and demand led to shifts in supply and demand curves and hence changed the location of the data points? Try drawing a diagram with a *given* demand curve and a *shifting* supply curve (do it now!). The equilibrium points you will trace out all lie on the *given* demand curve. If only supply shifts, we expect a *negative* relationship between price and quantity as we pick off different points on the same demand curve, which slopes downwards. Now, suppose the supply curve is *fixed* but the demand curve *shifts*. The equilibrium points then all lie on the *given* supply curve and exhibit a positive relation between price and quantity. The data in our example show a positive relationship between the price of houses and the quantity of new private houses built. Indeed, the two series of data move in a similar direction over time and hence principally correspond to a fixed supply curve for construction. It was demand for houses that must have been shifting around. House demand increased steadily during 1985-89, fell back during 1990-93, then grew again thereafter, until 2007 when the credit crunch took place.



UK house market 1985–2012 (2005 = 100)

Source: Nationwide.

Having made a diagnosis, we now gather corroborating evidence. Economy-wide activity is an important determinant of the demand for houses. UK real income grew strongly during 1985-89, fell sharply during 1990-93, grew fairly steadily thereafter and then fell again after 2007 because of the credit crunch. These changes in income fit nicely our theory that demand shifts are the main cause of the data pattern in the figure.

Summary

- **Demand** is the quantity that buyers wish to buy at each price. Other things equal, the lower the price, the higher the quantity demanded. Demand curves slope downwards.
- **Supply** is the quantity of a good sellers wish to sell at each price. Other things equal, the higher the price, the higher the quantity. Supply curves slope upwards.
- **The market clears**, or is in equilibrium, when the price equates the quantity supplied and the quantity demanded. At this point, supply and demand curves intersect. At prices below the equilibrium price there is **excess demand** (shortage), which itself tends to raise the price. At prices above the equilibrium price there is **excess supply** (surplus), which itself tends to reduce the price. In a **free market**, deviations from the equilibrium price tend to be self-correcting.
- Along a given demand curve, the other things assumed equal are the prices of related goods, consumer incomes and tastes or habits.
- An increase in the price of a **substitute** good (or decrease in the price of a **complementary** good) will raise the quantity demanded at each price. An increase in consumer income will increase demand for the good if the good is a **normal good** but decrease demand for the good if it is an **inferior good**.
- Along a given supply curve the other things assumed constant are technology, the price of inputs and the degree of government

regulation. An improvement in technology, or a reduction in input prices, will increase the quantity supplied at each price.

- Any factor inducing an increase in demand shifts the demand curve to the right, increasing equilibrium price and equilibrium quantity. A decrease in demand (downward shift of the demand curve) reduces both equilibrium price and equilibrium quantity. Any factor increasing supply shifts the supply curve to the right, increasing equilibrium quantity but reducing equilibrium price. Reductions in supply (leftward shift of the supply curve) reduce equilibrium quantity but increase equilibrium price.
- We can measure the **economic surplus** created by a market transaction by the sum of the consumer and producer surplus. The **consumer surplus** is measured by the area below the market demand and above the equilibrium price. The **producer surplus** is measured by the area above the market supply and below the equilibrium price.
- To be effective, a **price ceiling** must be imposed below the free market equilibrium price. It will then reduce the quantity supplied and lead to excess demand unless the government itself provides the extra quantity required. An effective **price floor** must be imposed above the free market equilibrium price. It will then reduce the quantity demanded unless the government adds its own demand to that of the private sector.

Review questions



EASY

1 The price and quantity data for toasters is given in the table below.

	Price					
Quantity	£10	£12	£14	£16	£18	£20
Demanded	10	9	8	7	6	5
Supplied	3	4	5	6	7	8

What is the excess supply or demand when the price is (a) £12 and (b) £20? Describe the price movements induced by positions (a) and (b).

- 2 Given that bread and toasters are complements, what happens to the demand curve for toasters if the price of bread rises? Show in a supply-demand diagram how the equilibrium price and quantity of toasters change.
- 3 How is the demand curve for toasters affected by the invention of the toaster oven if people prefer this new way of toasting? What happens to the equilibrium quantity and price of toasters?
- 4 You are a sheep farmer. Give three examples of a change that would reduce your supply of wool. Did you use a fall in the price of wool as one of your examples? Is it a valid example?
- 5 Goods with snob value are demanded because they are expensive. Does the demand curve for such goods slope upwards?

MEDIUM

- 6 The market demand for milk is $Q^D = 16 - 2P$, while the market supply is $Q^S = -2 + P$. Find the equilibrium quantity and price in the market for milk. Show your solution graphically.
- 7 Supply and demand data for toasters are shown below. Plot the supply curve and demand curve and find the equilibrium price and quantity.

	Price					
Quantity	£10	£12	£14	£16	£18	£20
Demanded	10	9	8	7	6	5
Supplied	3	4	5	6	7	8

- 8 Consider the following market demand $Q^D = 50 - 2P + Y$, where Y denotes consumers' income. Suppose that $Y = 10$ and then plot the market demand using the following price levels:

P					
3	5	7	9	12	15

Now suppose that consumers' income increases to $Y = 20$. Plot the market demand using the price levels in the table above. How has the rise in income affected the market demand?

- 9 Given the following market demand, $Q^D = 120 - 2P$, find the consumers' surplus when $P = £20$ and $P = £15$. Calculate a demand schedule and then draw a graph showing consumer surplus.
- 10 Given the following market supply, $Q^S = 210 - 5P$, find the producers' surplus when $P = 6$ and $P = 8$. Calculate a supply schedule and then draw a graph showing producer surplus.
- 11 Consider the market for safe cities. Someone knocks on your door and asks if you wish to purchase a reduction in crime by subscribing to an enhanced city-wide police force. Your city has 1 million residents. (a) What happens if you do not subscribe but all your fellow city dwellers do? (b) What happens if you subscribe but nobody else subscribes? (c) What does this tell you about the possibility of a market for public goods such as safe cities? (d) How might society ensure that desirably safe cities are provided?
- 12 Profitable speculation should stabilize financial markets - successful speculators are those who buy when the price is below the equilibrium price and sell when it has risen, or sell when the price is above the equilibrium price and buy when it has fallen. Why, then, are financial market prices so volatile?

HARD

- 13 **Essay question** The UK government is discussing a change in the planning laws to allow the building of 3 million new homes by 2020. Discuss what this is likely to mean for (a) the price of houses for first-time buyers and (b) the demand for country houses in areas adjacent to new housing developments. (c) Does your answer to (b) depend upon whether new houses are accompanied by new infrastructure (better roads, shops, train services, flood protection)?
- 14 The market data for butter are shown below.

Quantity	Price				
	£2	£7	£11	£13	£15
Demanded	105	80	60	50	40
Supplied	5	30	50	60	70

Suppose that the government introduces a price floor for butter at $P = £14$. (a) In a graph show the effect on the market for butter of such a price floor. (b) When will a price floor be inefficient? (c) What effect will a price floor have on consumers?

15 Common fallacies Why are these statements wrong? (a)

Manchester United is a more famous football club than Wrexham.
Manchester United will always find it easier to fill its stadium. (b)
Holding down rents ensures plenty of cheap housing for the poor.
(c) Stringent government regulations on safe production methods
benefit the consumers. (d) A legal minimum wage above the market
equilibrium makes all workers better off.

-
- 1 There are goods and services for which the law of demand does not hold. Nevertheless, those cases are rare. More details on the determinants of a demand curve are presented in Chapter 5, where the theory of consumer choice is presented.
 - 2 A price ceiling above the equilibrium price, such as P_2 in Figure 3.7, is irrelevant. The free market equilibrium at E is still to be attained. Suppliers have no incentive to charge a price higher than the equilibrium one. If they do that, they face lower demand and unsold stocks.

PART TWO

Positive microeconomics

Positive economics looks at how the economy functions. Microeconomics takes a detailed look at particular decisions without worrying about all the induced effects elsewhere. Part Two studies in detail the demand behaviour of consumers and the supply behaviour of producers, showing how markets work and why different markets exhibit different forms of behaviour. By applying similar tools to the analysis of input markets, we can also understand why some people earn so much more than others.

Chapter 4 examines the responsiveness of demand and supply behaviour. Chapter 5 develops a theory of demand based on self-interested choice by consumers. Chapter 6 introduces different types of firm and considers motives behind production decisions. Chapter 7 analyses how costs of production influence the output that firms choose to supply. Chapter 8 and Chapter 9 explore how differences in market structure affect competition and the output decision of firms. Chapter 10 and Chapter 11 analyse input markets for labour, capital and land, which determine the distribution of income. Chapter 12 explains why people dislike risk, how institutions develop to shift risk on to those who can bear it more cheaply, and why informational problems can inhibit the development of markets for some commodities.

Contents

1. Elasticities of demand and supply
2. Consumer choice and demand decisions
3. Introducing supply decisions
4. Costs and supply
5. Perfect competition and pure monopoly
6. Market structure and imperfect competition
7. The labour market
8. Factor markets and income distribution

9. Risk and information

CHAPTER 4

Elasticities of demand and supply

Learning outcomes

By the end of this chapter, you should be able to:

- 1 describe how elasticities measure responsiveness of demand or supply
- 2 define the price elasticity of demand
- 3 understand how the price elasticity of demand affects the revenue effect of a price change
- 4 realize why bad harvests may help farmers
- 5 recognize the fallacy of composition
- 6 understand how cross-price elasticity relates to complements and substitutes
- 7 describe income elasticity of demand
- 8 recognize inferior, normal and luxury goods
- 9 analyse elasticity of supply
- 10 understand how supply and demand elasticities affect tax incidence

In Chapter 3 we examined how the price of a good affects the quantity demanded. We saw that changes in income, or in the price of related goods, shift demand curves, altering the quantity demanded at each price. We now study these effects in more detail.

4.1

The price responsiveness of demand

A downward-sloping demand curve shows that lower prices increase quantity demanded. Often, we need to know by how much quantity will increase. Table 4.1 presents some hypothetical numbers relating ticket price and quantity demanded, other things equal. From columns (1) and (2), Figure 4.1 plots the demand curve, which happens to be a straight line.

How do we measure the responsiveness of the quantity of tickets demanded to the price of tickets? An obvious measure is the slope of the demand curve. Each price cut of £1 leads to 8000 extra ticket sales.¹

Suppose we want to compare the price responsiveness of football ticket sales with that of cars. Using only the slopes of the demand curves makes this comparison not particularly attractive. £1 is a trivial cut in the price of a car and has a negligible effect on the quantity of cars demanded. We need a way to normalize the slope of the demand function in order to make useful comparisons among different goods.

Table 4.1 The demand for football tickets

(1) Price (£/ticket)	(2) Tickets demanded ('000s)	(3) Price elasticity of demand
12.50	0	$-\infty$
10.00	20	-4
7.50	40	-1.5
5.00	60	-0.67
2.50	80	-0.25
0	100	0

The **price elasticity of demand (PED)** is the percentage change in the quantity demanded divided by the corresponding percentage change in its price.

$$\text{PED} = (\% \text{ change in quantity}) / (\% \text{ change in price})$$

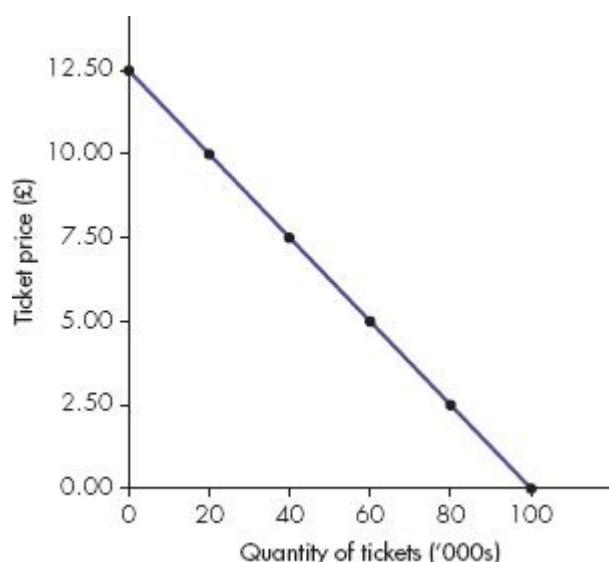
In Chapter 2 we argued that, when commodities are measured in different units, it is often best to examine the percentage change, which is unit-free.

Although we later introduce other demand elasticities – the cross-price and the income elasticities – the (own-)price elasticity is the most often

used of the three. If economists speak of the demand elasticity, they mean the **price elasticity of demand (PED)**.

Suppose a 1 per cent price rise reduces the quantity demanded by 2 per cent. The demand elasticity is the percentage change in quantity (-2) divided by the percentage change in price ($+1$) and is thus given by -2 . The minus sign tells us that quantity *falls* when price rises. If a price fall of 4 per cent increases the quantity demanded by 2 per cent, the demand elasticity is $-\frac{1}{2}$ since the quantity change ($+2$ per cent) is divided by the price change (-4 per cent). Since demand curves slope down, price and quantity changes always have opposite signs. The price elasticity of demand tells us about movements along a demand curve. The demand elasticity is a negative number.

For further brevity, economists often omit the minus sign. It is easier to say the demand elasticity is 2 than to say it is -2 . When the price elasticity of demand is expressed as a positive number, it is implicit that a minus sign must be added (unless there is an explicit warning to the contrary). Otherwise, it implies that demand curves slope up, a rare but not unknown phenomenon.



For given prices of related goods and consumer incomes, higher ticket prices reduce the quantity of tickets demanded.

Figure 4.1 The demand for football tickets

The price elasticity of demand for football tickets is shown in column (3) of Table 4.1. Examining the effect of price cuts of £2.50, we calculate the price elasticity of demand at each price. Beginning at £10 and 20 000 tickets demanded, consider a price cut to £7.50. The price change is -25 per cent, from £10 to £7.50; the change in quantity demanded is $+100$ per cent, from 20000 to 40000 tickets.

The demand elasticity at £10 is $(100/-25) = -4$. Other elasticities are calculated in the same way, dividing the percentage change in quantity by the corresponding percentage change in price. When we begin from the price of £12.50, the demand elasticity is minus infinity. The percentage change in quantity demanded is $+20/0$. Any positive number divided by zero tends to infinity. Dividing by the -20 per cent change in price, from £12.50 to £10, the demand elasticity tends to minus infinity at this price.

$$\text{Price elasticity of demand} = [\% \text{ change in quantity}] / [\% \text{ change in price}]$$

We say that the demand elasticity is *high* if it is a large negative number. The quantity demanded is sensitive to the price. The demand elasticity is *low* if it is a small negative number and the quantity demanded is insensitive to the price. ‘High’ or ‘low’ refer to the size of the elasticity, ignoring the minus sign. The demand elasticity falls when it becomes a smaller negative number and quantity demanded becomes less sensitive to the price.²

MATHS 4.1

PRACTISING CALCULATION OF PRICE ELASTICITY OF DEMAND (PED) AND THE ARC ELASTICITY OF DEMAND

P = price (£)	1	2	3	4	5	6
Q ½ quantity demanded	10	8	6	4	2	1

The rows in the table above give price and quantity data for a particular demand curve. The table below shows five columns, labelled A-E, each corresponding to a situation in which the price changes by £1 and there is a corresponding change in the quantity demanded.

In column A, a 100 per cent price rise (from £1 to £2) induces a 20 per cent fall in quantity demanded (from 10 to 8), implying a price elasticity of demand of $(-20/100) = -0.2$. Similarly, in column C, a 50 per cent price reduction (from £2 to £1) induces a 25 per cent rise in quantity demanded (from 8 to 10), implying a price elasticity of $(25)/(-50) = -0.5$.

	A	B	C	D	E
(1) Initial P and Q	$P = 1$ $Q = 10$	$P = 2$ $Q = 8$	$P = 2$ $Q = 8$	$P = 4$ $Q = 4$	$P = 5$ $Q = 2$
(2) New P and Q	$P = 2$ $Q = 8$	$P = 3$ $Q = 6$	$P = 1$ $Q = 10$	$P = 3$ $Q = 6$	$P = 6$ $Q = 1$
(3) % change in P	$100 \times 2 - 1 /1 = 100$		$100 \times 1 - 2 /2 = -50$		
(4) % change in Q thus induced	$100 \times 8 - 10 /10 = -20$		$100 \times 10 - 8 /8 = 25$		
(5) $PED = (4)/(3)$	-0.2	-0.5			

Notice the asymmetry in the calculation of the elasticity of demand using our definition. If we consider an increase in the price from £1 to £2 (and a corresponding decrease in quantity from 10 to 8), we obtain a given value for the PED (-0.2). If we consider the opposite case (a decrease in the price from £2 to £1 and a corresponding increase in the quantity from 8 to 10), we obtain a different value for the PED (-0.5). To avoid this asymmetry, it is possible to create a refinement in the calculation of the elasticity of demand. We can calculate the arc elasticity of demand. The arc elasticity of demand can be calculated using the following formula:

$$PED_{arc} = \frac{(Q_1 - Q_0)}{(Q_1 + Q_0)/2} \Big/ \frac{(P_1 - P_0)}{(P_1 + P_0)/2}$$

where Q_1 and P_1 are the new values of quantity and price, respectively, while Q_0 and P_0 are the initial values. In practice, we express the change in price as a percentage of the average price minus the midpoint between the initial and new price. Similarly, we express the change in the quantity demanded as a percentage of the average quantity demanded minus the average of the initial and new quantity. The advantage of using the average price and the average quantity is that the value of the elasticity is the same whether the price rises or falls.

Consider the case in column A: $P_0 = 1$, $Q_0 = 10$, $P_1 = 2$, $Q_1 = 8$. The arc elasticity of demand in this case is:

$$PED_{arc} = \frac{(8 - 10)}{(8 + 10)/2} / \frac{(2 - 1)}{(2 + 1)/2} = -0.148$$

Now consider the case in column C: $P_0 = 2$, $Q_0 = 8$, $P_1 = 1$, $Q_1 = 10$. The arc elasticity of demand is:

$$PED_{arc} = \frac{(10 - 8)}{(10 + 8)/2} / \frac{(1 - 2)}{(1 + 2)/2} = -0.148$$

Using the arc elasticity of demand the result is the same whether we consider a price cut of a given amount or a price increase by the same amount. This is true regardless of whether the demand is linear or non-linear.

Which method should we use to calculate the elasticity of demand? In general, it is better to use the arc elasticity of demand if the form of the demand is unknown.

Applying the normal definition of the *PED* or the arc elasticity version to a given set of observations on prices and quantities may lead to different numerical results. Nevertheless, in general, the main properties of the elasticity (meaning if a demand is elastic or inelastic between two different points) would be similar independently of the way we calculate the elasticity.

Questions

- a. Try to complete columns B, D and E for yourself.
- b. Complete columns B, D and E using the arc elasticity of demand.

To check your answers to these questions, go to page 677.

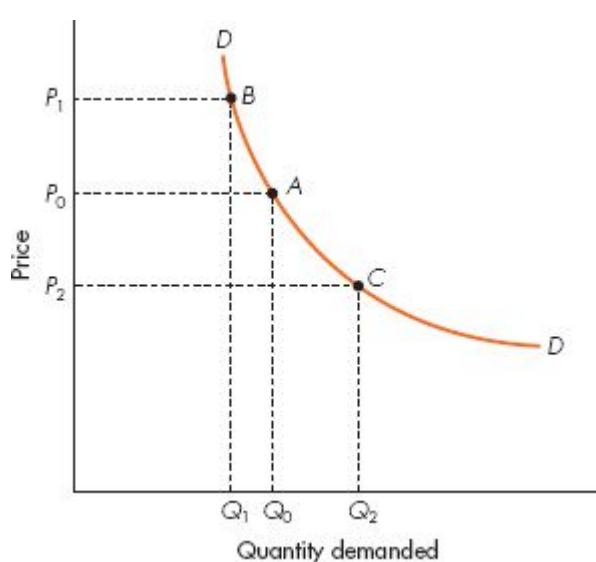
The demand curve for football tickets is a straight line with a constant slope: along its entire length a £1 cut in price always leads to 8000 extra ticket sales. Yet Table 4.1 shows that demand elasticity falls as we move down the demand curve from higher prices to lower prices. At high prices, £1 is a small percentage change in the price but 8000 tickets is a large percentage change in the quantity demanded. Conversely, at low

prices £1 is a large percentage change in the price but 8000 is a small percentage change in the quantity. When the demand curve is a straight line, the price elasticity falls steadily as we move down the demand curve.³

It is possible to construct curved demand schedules (still, of course, sloping downwards) along which the price elasticity of demand remains constant. Generally, however, the price elasticity changes as we move along demand curves, and we expect the elasticity to be high at high prices and low at low prices.

If the demand curve is a straight line, we get the same size of quantity response (20 000 tickets) whether we raise or lower the price by £2.50. It does not matter whether we use price rises or price cuts to calculate the demand elasticity.

When, as in Figure 4.2, the demand curve is not a straight line, we meet a minor difficulty. Beginning at point *A* where the price is P_0 , moves to points *B* and *C* are percentage price changes of equal magnitude but opposite sign. Figure 4.2 shows that the quantity response (from Q_0 to either Q_1 or Q_2) differs for price rises and price falls when the demand curve is not a straight line.



When the demand curve *DD* is non-linear, price rises and price cuts of equal size lead to quantity changes that differ in size.

Figure 4.2 A non-linear demand curve *DD*

For non-linear demand curves, economists resolve this ambiguity about the definition of price elasticity of demand by defining it with respect to *very small* changes in price (see Maths 4.2). If we move only a short distance either side of point A, the demand curve hardly has time to bend round. Over the very short distance corresponding to a small percentage price rise or fall, the demand curve is as near a straight line as makes no difference. With this amendment we can use the old definition.

Elastic and inelastic demand

Although elasticity typically falls as we move down the demand curve, an important dividing line occurs at the demand elasticity of -1 . In Table 4.1 demand is **elastic** at all prices of £7.50 and above and **inelastic** at all prices of £5 and below.

Demand is **elastic** if price elasticity is more negative than -1 .

Demand is **inelastic** if the price elasticity lies between -1 and 0 .

Although the price elasticity of demand typically changes as we move along demand curves, economists frequently talk of goods with high or low demand elasticities. For example, the demand for oil is price inelastic (price changes have only a small effect on quantity demanded) but the demand for foreign holidays is price elastic (price changes have a big effect on quantity demanded). Such statements implicitly refer to parts of the demand curve corresponding to prices usually charged for these goods or services.

MATHS 4.2

THE POINT ELASTICITY OF DEMAND

We have defined the price elasticity of demand as:

$$PED = \frac{\% \Delta Q^D}{\% \Delta P} \quad (1)$$

where the Greek letter delta (δ) stands for ‘change’ and $\% \delta Q^D$ denotes the percentage change in the quantity demanded and $\% \delta P$ denotes the percentage change in the price. Mathematically, the percentage change in the quantity demanded can be written as:

$$\% \Delta Q^D = \frac{\Delta Q^D}{Q^D} \times 100$$

Similarly, the percentage change in the price can be written as:

$$\% \Delta P = \frac{\Delta P}{P} \times 100$$

Using those facts in relation to the definition of *PED* in equation (1) and after some algebra we have:

$$PED = \frac{\Delta Q^D}{\Delta P} \frac{P}{Q^D} \quad (2)$$

Equation (2) is another way to write the price elasticity of demand. We can use the formula in (1) or the idea of arc elasticity of demand to find the elasticity between two different points on a demand curve. However, in many cases we are interested in calculating the elasticity of demand at a given point of a specific demand function.

Suppose we face the following linear direct demand function for a given good:

$$Q^D = 100 - 2P \quad (3)$$

where Q^D is the quantity demanded and P is the price.

When the price is $P = 10$, according to the demand in equation (3), the quantity demanded is $Q^D = 80$. Is the demand elastic, or inelastic, at $P = 10$?

To answer that question we use the concept of point elasticity of demand, since we want to measure the elasticity at a particular point of a demand function.

The point elasticity of demand is defined as:

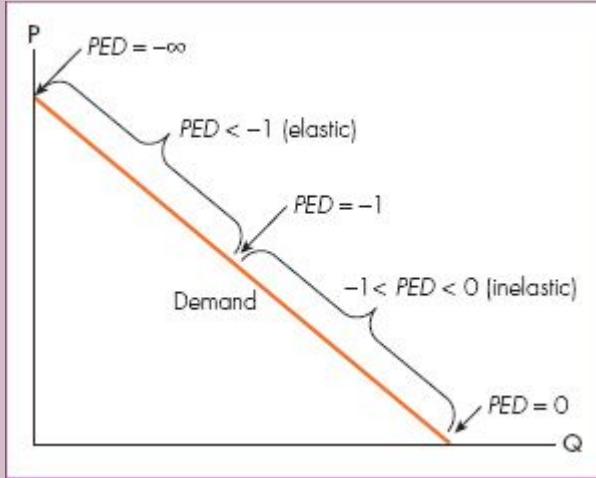
$$PED = \frac{dQ^D}{dP} \frac{P}{Q^D} \quad (4)$$

where dP is now a very small (close to zero) change in the price and dQ^D is the corresponding change in the quantity demanded. The term dQ^D/dP is the *derivative* of the direct demand function with respect to the price. It measures the slope of the direct demand function at a given point. In the case of a linear direct demand

function, the slope is constant along the demand curve. In our case, the slope of the direct demand curve is $dQ^D/dP = -2$.

Using equation (4), we have that at $P = 10$ and $Q^D = 80$ the point elasticity of demand is:

$$PED = -2 \frac{10}{80} = -0.25$$



The demand is therefore inelastic when $P = 10$.

From equation (4) we can see why, along a linear demand curve, the elasticity of demand is not constant. While the slope is constant, the term P/Q changes along the demand curve. This is summarized in the figure on the right.

Using equation (4) you can find the elasticity at any point on a demand curve, linear or nonlinear. There are special cases in which the elasticity is constant along a demand curve. Consider the following direct demand function:

$$Q^D = \frac{100}{P} \quad (5)$$

The demand function in (5) is non-linear (try to plot it!). In this case, the derivative of the direct demand is:

$$\frac{dQ^D}{dP} = \frac{100}{P^2}$$

Suppose we want to know the elasticity of demand at $P = 5$. At that price the quantity demanded according to (5) is $Q^D = 20$. In this

case, we have $dQ^D/dp = -[100/(5)^2] = -4$. Applying equation (4), we have:

$$PED = -4 \frac{5}{20} = -1$$

At $P = 5$, the demand is **unit elastic**. What about at $P = 10$? In this case, the quantity demanded is $Q^D = 100/10 = 10$ and then $dQ^D/dp = -[100/(10)^2] = -1$. The PED is:

If the demand elasticity is -1 , demand is **unit elastic**.

$$PED = -1 \frac{10}{10} = -1$$

Thus, the PED is still -1 . Indeed, any demand function of the form $Q^D = A/P$, where A is any positive constant, has the property that the elasticity of demand is constant along the demand curve.

Determinants of price elasticity

Why is the price elasticity of demand for a good high (-5) or low (-0.5)? The answer lies in consumer tastes. If it is a social necessity to own a television, higher TV prices have little effect on quantity demanded. If TVs are considered a frivolous luxury, the demand elasticity is much higher. Psychologists and sociologists can explain why tastes are as they are. As economists, we can identify some considerations likely to affect consumer responses to changes in the price of a good. *The most important consideration is the ease with which consumers can substitute another good that fulfils approximately the same function.*

Consider two extreme cases. Suppose the price of all cigarettes rises by 1 per cent. The quantity of cigarettes demanded will hardly respond. People who can easily quit smoking have already done so. In contrast, suppose the price of a particular brand of cigarettes rises by 1 per cent, all other brand prices remaining unchanged. We expect a much larger quantity response. Consumers switch from the dearer brand to other

brands that also satisfy the nicotine habit. For a particular cigarette brand the demand elasticity is quite high.

Our example suggests a general rule. The more narrowly we define a commodity (a particular brand of cigarette rather than cigarettes in general), the higher will be the price elasticity of demand.

Measuring price elasticities

Table 4.2 confirms that the demand for broad categories of basic commodities, such as fuel, food or even services to households, is inelastic. As a category, only alcohol seems to have an elastic demand. Households simply do not have much scope to alter the broad pattern of their purchases.

Table 4.2 UK price elasticities of demand

Good (broad type)	Demand elasticity	Good (narrow type)	Demand elasticity
Fuel and light	-0.52	Bread	-0.4
Food	-0.56	Fish	-0.8
Clothing	-0.62	Beer	-0.2
Services	-0.72	Expenditure abroad	-1.6
Alcohol	-1.73	Catering	-2.6

Source: R. Blundell, P. Pashardes and G. Weber, ‘What do we learn about consumer demand patterns from micro data?’, American Economic Review 83, no. 3 (1993): 570–597; National Food Survey 2000.

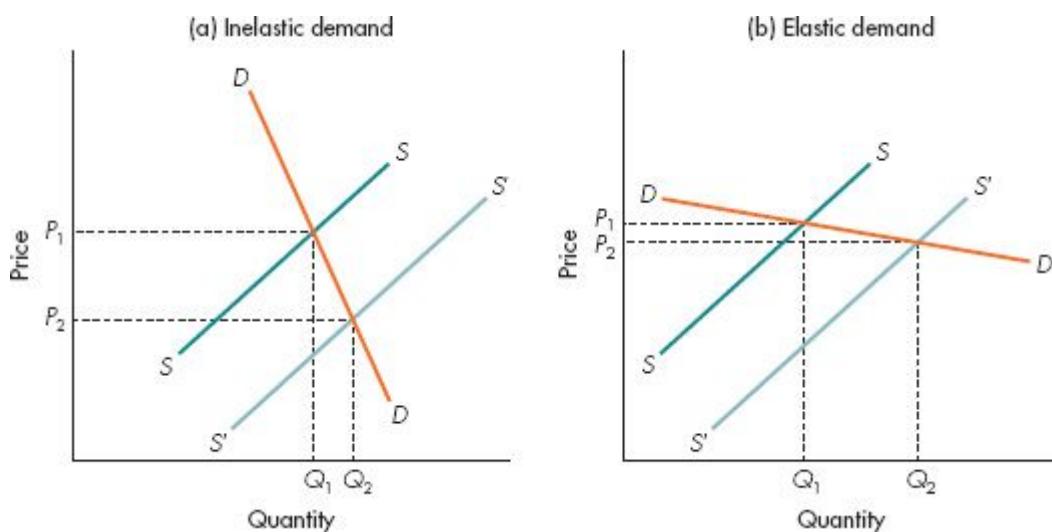
In contrast, there is a much wider variation in the demand elasticities for narrower definitions of commodities. Even then, the demand for some commodities, such as dairy produce, is very inelastic. However, particular kinds of services such as catering have much more elastic demand.

Using price elasticities

Price elasticities of demand are useful in calculating the price rise required to eliminate a shortage (excess demand) or the price fall to eliminate a surplus (excess supply). One important source of surpluses and shortages is shifts in the supply curve. Harvest failures (and bumper crops) are a feature of agricultural markets. Because the demand

elasticity for many agricultural products is very low, harvest failures produce large increases in the price of food. Conversely, bumper crops induce very large falls in food prices. When demand is very inelastic, shifts in the supply curve lead to large fluctuations in price but have little effect on equilibrium quantities.

Figure 4.3(a) illustrates this. SS is the supply curve in an agricultural market when there is a harvest failure and $S'S'$ is the supply curve when there is a bumper crop. The equilibrium price fluctuates between P_1 (harvest failure) and P_2 (bumper crop) but induces little fluctuation in the corresponding equilibrium quantities. Contrast this with Figure 4.3(b), which shows the effect of similar supply shifts in a market with very elastic demand. Price fluctuations are much smaller but quantity fluctuations are now much larger. Knowing the demand elasticity helps us understand why some markets exhibit volatile quantities but stable prices, while other markets exhibit volatile prices but stable quantities.



In each case, the supply curve fluctuates between SS and $S'S'$. In case (a), demand is inelastic, and supply shifts lead to large changes in equilibrium price but little change in equilibrium quantity. In case (b), demand is elastic, and the same supply shift now leads to large changes in equilibrium quantity but little change in equilibrium price.

Figure 4.3 The effect of demand elasticity on equilibrium price and quantity fluctuations

4.2 Price, quantity demanded and total expenditure

Other things equal, the demand curve shows how much consumers of a good wish to purchase at each price. At each price, total spending by consumers is the price multiplied by the quantity demanded. This represents the total revenues received by suppliers. We now discuss the relationship between total spending and price and show the relevance of the price elasticity of demand.

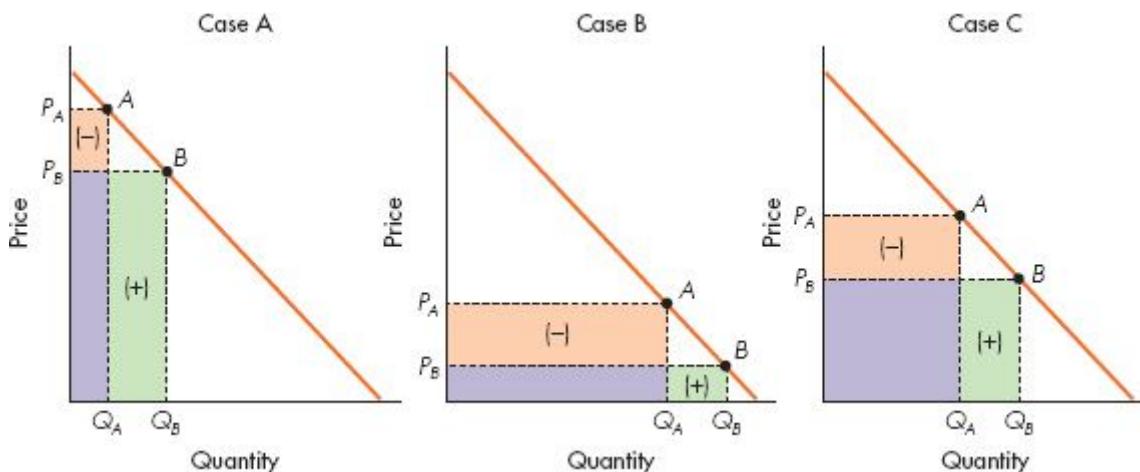
Consider a price cut in the market. This creates two effects:

- (1) *A price effect*: each unit of the good is now sold at a lower price, which tends to decrease total spending.
- (2) *A quantity effect*: after the price cut, more units are sold, which tends to raise total spending.

The elasticity of demand tells us which of the two effects dominates.

Figure 4.4 shows how total spending changes with price changes. In case A, we begin at A with price P_A and quantity demanded Q_A . Total spending is $P_A Q_A$, the area of the rectangle $OP_A A Q_A$. At the lower price P_B , consumers demand Q_B and total spending is $P_B Q_B$, the area of the rectangle $OP_B B Q_B$. How does total spending change when prices fall from P_A to P_B ? Spending falls by the area marked $(-)$ but rises by the area marked $(+)$. In case A, the $(+)$ area exceeds the $(-)$ area and total spending rises. In the elastic range of the demand curve (towards the upper end), a lower price raises the quantity demanded by more than enough to offset the lower price. The quantity effect dominates and total spending rises.

Case B examines the lower end of the demand curve, where demand is inelastic. Although the price cut raises the quantity demanded, the rise in quantity is insufficient to compensate for the lower price. The $(+)$ area is smaller than the $(-)$ area. The price effect dominates and total spending falls. If price cuts increase total spending at high prices where the demand elasticity is high and reduce total spending at low prices where the demand elasticity is low, at some intermediate price a fall in price will leave total spending unaltered. Case C shows this possibility. The higher quantity demanded exactly compensates for the lower price.



When the price is reduced from P_A to P_B , expenditure changes from $OP_A A Q_A$ to $OP_B B Q_B$. Spending rises when demand is elastic (case A), falls when demand is inelastic (case B), and is unchanged when demand is unit elastic (case C).

Figure 4.4 Elasticity of demand and the effect of price changes on expenditure

If quantity demanded rises 1 per cent when the price falls 1 per cent, the price effect and the quantity effect exactly balance and total spending is unchanged. Case C shows the point on the demand curve at which the price elasticity of demand is -1 (quantity change -1 per cent; price change -1 per cent). If demand is elastic, as in case A, a 1 per cent price cut leads to an increase in quantity by *more* than 1 per cent. Hence total spending rises. Conversely, when demand is inelastic, as in case B, a 1 per cent price cut leads to an increase in quantity by *less* than 1 per cent. Hence total spending falls. These results are summarized in Table 4.3.

Table 4.3 Demand elasticities and changes in spending

Change in total spending caused by	Price elasticity of demand		
	Elastic (e.g. -3)	Unit elastic (-1)	Inelastic (e.g. -0.3)
Price rise	Fall	Unchanged	Rise
Price cut	Rise	Unchanged	Fall

CASE 4.1

THE PRICE OF THE IPHONE AND THE ELASTICITY OF DEMAND

The iPhone 8GB, a multimedia smart phone created by Apple, was launched in the US market in June 2007. The launch price was \$599. In early September of the same year Apple announced a reduction in the price of the iPhone 8GB of 33 per cent, from \$599 to \$399. Why such a big reduction in the price a mere two months after the launch?

One possibility is that Apple overestimated its demand for the iPhone and not many customers were willing to buy at that price. In this case, a price cut may be justified. This possibility, however, seems implausible. Demand for iPhones was high; customers were queuing outside Apple stores for hours in order to buy one. Maybe the price cut was due to the fact that Apple discovered that the cost of making the iPhone was lower than expected. This possibility also seems quite implausible. According to the market research company iSupply, the cost of making the iPhone 8GB was \$280.83 when the iPhone was launched. This was the cost when the price was \$599 and also when the price was reduced.

A more plausible explanation is related to the idea of the elasticity of demand. The iPhone can be viewed as a luxury good with few direct substitutes. This would imply that the price elasticity of demand for the iPhone should probably be low. In pricing the iPhone, Apple would like to set a price such that the elasticity of demand is close to -1 . It turned out that this was not the case. Indeed, it seems that the demand for the iPhone was more elastic than that. Various estimates found an elasticity of demand between -3 and -5 per cent. According to our analysis, if the initial price is on the highly elastic part of the demand curve, a reduction in the price increases total expenditure, that is, price multiplied by quantity. But total expenditure represents the total revenue received by the firm that is selling the good. So a decrease in price when demand is quite elastic will increase the revenues obtained by the firm. To gain a rough idea of the elasticity of demand for the iPhone we can use the sales data from Apple. In the first three months after the iPhone was launched, Apple sold 270 000 iPhones and the price was mainly \$599. In the fourth quarter of 2007 (from October to

December), Apple sold 1 119 000 iPhones when the price was \$399. Using the arc elasticity of demand equation with those data, the elasticity of demand is approximately -3.06 per cent. While this is a rough measure, it still gives us an idea of how elastic the demand for iPhones is. According to this rough measure, a decrease in the price by 1 per cent will increase quantity demanded (and so sold) by 3 per cent.

Obviously the cut in the price made the customers who bought the iPhone at \$599 quite unhappy. A \$100 discount voucher to be spent in Apple stores was given to those customers to partially compensate them.

Source: Compiled by the authors.

The price of football tickets and total spending

Think again about revenue from ticket sales. Table 4.4 shows the demand data of Table 4.1, but also shows the tickets demanded at a price of £6.25 per ticket. At this price the demand elasticity is -1 . A 20 per cent price cut ($-\text{£}1.25$) induces a 20 per cent rise in the quantity demanded (10 000 tickets). Column (4) shows total spending on tickets at each price.

Table 4.4 Ticket demand and revenue

(1) Ticket price (£)	(2) Quantity demanded ('000s)	(3) Price elasticity of demand	(4) Total spending (£000s)
12.50	0	$-\infty$	0
10.00	20	-4	200
7.50	40	-1.5	300
6.25	50	-1	312.5
5.00	60	-0.67	300
2.50	80	-0.25	200
0	100	0	0

Beginning from the highest price of £12.50, successive price cuts first increase total spending on tickets, then reduce it. Table 4.4 explains why. When the price is high, demand is elastic: price reductions increase total spending. When demand is unit elastic, at price £6.25, we reach a turning

point. Above this price, price cuts have steadily increased total spending. Below this price, further price cuts reduce total spending because demand is then inelastic.

We can thus draw the following conclusion: *spending and revenue reach a maximum at the point of unit-elastic demand*. This idea, and the empirical knowledge that this occurs at the price of £6.25 per ticket, are the pieces of information the football club owner needs to know.

4.3

Further applications of the price elasticity of demand

The coffee frost

There's an awful lot of coffee in Brazil— the country supplies a large share of the world market. In 1994 people first realized that a frost in Brazil would cause havoc with the 1995 harvest. *The Economist* reported estimates that the 1995 crop would not be the 26.5 million bags previously thought, but only 15.7 million bags.⁴ Obviously, coffee was going to be scarce in 1995. Anticipating this, speculators bought coffee in 1994, bidding up its price even before the supply fell.

Table 4.5 shows the effect on Brazilian exports during 1993–95. The first row shows that, even after adjusting for general inflation, coffee prices more than doubled in US dollars. The second row shows an index of the volume of Brazilian coffee exports. The final row shows Brazilian export revenue from coffee. Real revenue rose sharply in 1994: prices had risen *before* production had fallen too much. The interesting comparison is between 1993 and 1995. Brazilian export revenue from coffee *increased* despite the 'bad' harvest.

Table 4.5 Brazilian coffee exports

	1993	1994	1995
Price (US\$/lb)	0.9	2.0	2.1
Export quantity (1990 = 100)	113	102	85
Price X quantity	102	204	179

Note: Prices are in 1995 US\$.

Sources: IMF, *International Financial Statistics*.

The demand for coffee is inelastic, despite an abundance of substitutes – tea, soft drinks and beer. This example emphasizes the importance of consumer taste. If buyers refuse to abandon coffee drinking, it is useless to point out that a blend of tea and Coca-Cola contains as much caffeine as the average cup of coffee.

Farmers and bad harvests

This example illustrates a general result. When demand is inelastic farmers earn more revenue from a bad harvest than from a good one. When the supply curve shifts to the left, it takes a big rise in price to eliminate excess demand when demand is inelastic. And price increases raise consumer spending and producer revenues when demand is inelastic. Demand elasticities are low for many commodities such as coffee, milk and wheat. They are part of our staple diet. Eating habits are slow to change, even when prices rise.

If bad harvests raise farmers' revenues and good harvests lead to a fall in agricultural prices and farmers' revenues, why don't farmers get together like OPEC to restrict their supply and increase revenues in the face of inelastic demand? If it were easy to organize such collusion between farmers, it would occur more frequently. Later we discuss the difficulties that arise in trying to maintain a co-operative policy to restrict supply.

When demand is inelastic, suppliers taken together are better off if supply can be reduced. However, if one farmer loses part of the crop but all other farmers's crops are unaffected, the unlucky farmer is worse off – the **fallacy of composition**. The fall in a single farmer's output, unlike the reduction of all farmers's outputs simultaneously, has a negligible effect on supply. Market price is unaffected and the unlucky farmer simply sells less output at the price that would have prevailed in any case. This illustrates an important lesson in economics. The individual producer faces a demand that is very elastic – consumers can easily switch to the output of similar farmers – even if the demand for the crop as a whole is very inelastic.

The **fallacy of composition** means that what is true for the individual may not be true for everyone together, and what is true for everyone together may not hold for the individual.

CASE 4.2

EASY PROFITS

Low-cost airline pioneer Sir Stelios Haji-Ioannou, founder of easyJet and then the EasyGroup, credits two things with his success. The first, which he says only half in jest, is coming from a rich family, which made it easier to get through the early years. The second, which he also proudly cites, is his economics degree, where the lecture on elasticity of demand helped underpin his conviction that low prices could generate large revenues by creating high sales volume. A recent estimate of the fare elasticity of demand for air travel in Europe is -1.40 , meaning that the demand is indeed elastic.⁵ When he launched easyJet, conventional airlines were happy to fill 70 per cent of their seats on an average flight. EasyJet now runs regularly at 85 per cent capacity on its 600 flights a day, which is a lot of extra revenue without any additional costs.



© Ice962 | Dreamstime.com
Source: www.easyJet.com.

4.4 Short run and long run

The price elasticity of demand varies according to the length of time in which consumers can adjust their spending patterns when prices change. The most dramatic price change of the past 50 years, the oil price rise of 1973–74, caught many households with a new but fuel-inefficient car. At first, they may not have expected the higher oil price to last. Then they may have planned to buy a smaller car with greater fuel efficiency. But in countries like the US, few small cars were yet available. In the short

run, households were stuck. Unless they could rearrange their lifestyles to reduce car use, they had to pay the higher petrol prices. Demand for petrol was inelastic.

Over the long run, consumers had time to sell their big cars and buy cars with better fuel efficiency, or to move from the distant suburbs closer to their place of work. Over this longer period, they could reduce the quantity of petrol demanded much more than they could initially.

The price elasticity of demand is lower in the short run than in the long run when there is more scope to substitute other goods. This result is very general. Even if addicted smokers cannot adjust to a rise in the price of cigarettes, fewer young people start smoking and gradually the number of smokers falls.

Table 4.6 reports estimates of the short- and long-run elasticities of demand for various goods in the UK. Those results show that the long-run elasticity tends to be larger than the short-run one.

The **short run** is the period after prices change but before quantity adjustment can occur.

The **long run** is the period needed for complete adjustment to a price change. Its length depends on the type of adjustments consumers wish to make.

Table 4.6 Short- and long-run elasticities in the UK

	Short run	Long run
Bus service	-0.43	- 1.25
Underground service	-0.31	- 0.57
Mobile calls	-0.78	- 1.04

Sources: A. Alpetkin et al., ‘Estimating spectrum demand for the cellular services in the UK’, working paper, University of Surrey, 2007; M. Wardman and J. Shires, ‘Review of fare elasticities in Great Britain’, ITS working paper, 2003.

How long is the long run?

There is no definite answer to this question. Demand responses to a change in the price of chocolate for example should be completed within

a few months, but full adjustment to changes in the price of oil or cigarettes may take years.

4.5 The cross-price elasticity of demand

The price elasticity of demand tells us about movements along a given demand curve, holding constant all determinants of demand except the price of the good itself. We now hold constant the own-price of the good and examine changes in the prices of related goods. The cross-price elasticity tells us the effect on the quantity demanded of good i when the price of good j is changed. As before, we use percentage changes.

The cross-price elasticity may be positive or negative. It is positive if a rise in the price of good j increases the quantity demanded of good i . Suppose good i is Coke and good j is Pepsi. An increase in the price of Coke raises the demand for Pepsi. The cross-price elasticity of Coke with respect to Pepsi is positive. Cross-price elasticities are positive when two goods are substitutes and negative when two goods are complements.⁶ We expect a rise in the price of petrol to reduce the demand for cars because petrol and cars are complements.

If the cross-price elasticity between two goods is zero or close to zero, we say the two goods are independent. We do not expect that a change in the price of petrol will affect the demand for ice cream.

The **cross-price elasticity of demand** for good i with respect to changes in the price of good j is the percentage change in the quantity of good i demanded, divided by the corresponding percentage change in the price of good j .

Table 4.7 shows estimates for the UK for three different goods. Own-price elasticities for food, clothing and travel are given diagonally down the table, from top left (the own-price elasticity of demand for food) to bottom right (the own-price elasticity of demand for travel). Off-diagonal entries in the table show cross-price elasticities of demand. Thus, 0.1 is the cross-price elasticity of demand for food with respect to transport. A 1 per cent increase in the price of travel increases the quantity of food demanded by 0.1 per cent.

Table 4.7 Cross-price and own-price elasticities of demand in the UK

% change in quantity	Caused by a 1% price change in demand for		
	Food	Clothing	Travel
Food	-0.4	0	0.1
Clothing	0.1	-0.5	-0.1
Travel	0.3	-0.1	-0.5

Sources: R. Blundell et al., 'What do we learn about consumer demand patterns from micro data?', *American Economic Review* 83, no. 3 (1993): 570–597.

The own-price elasticities for the three goods lie between -0.4 and -0.5. For all three goods, the quantity demanded is more sensitive to changes in its own price than to changes in the price of any other good.

4.6 The effect of income on demand

Finally, holding constant the own-price of a good and the prices of related goods, we examine the response of the quantity demanded to changes in consumer incomes. For the moment, we neglect the possibility of saving. Thus a rise in consumer income will typically be matched by an equivalent increase in total consumer spending.

Chapter 3 pointed out that higher consumer incomes tend to increase the quantity demanded. However, demand quantities increase by different amounts as incomes rise. Thus the pattern of consumer spending on different goods depends on the level of consumer income. The **budget share** of a good is the fraction of total consumer spending for which it accounts.

Table 4.8 reports the share of consumer spending in the UK devoted to food and drink and to recreation and cultural goods between 1997 and 2005. Real consumer spending (and incomes) rose during 1997–2005. Even though real spending on food and drink increased, its budget share fell. Spending on recreation and cultural goods rose so much that its budget share increased substantially. These changes in budget share mainly reflect changes in real consumer incomes and different **income elasticities of demand**.

The **budget share** of a good is its price multiplied by the quantity demanded, divided by total consumer spending or income.

The **income elasticity of demand** for a good is the percentage change in quantity demanded divided by the corresponding percentage change in consumer income.

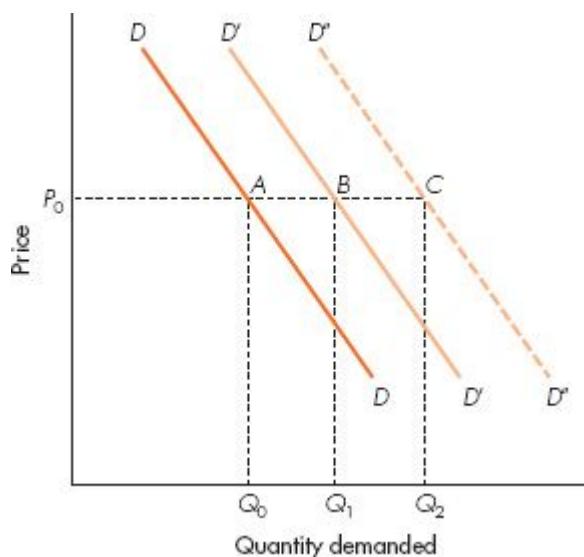
Table 4.8 Budget shares, 1997-2005

	Real consumer spending (2003 £bn)	% budget share	
		Food and drink	Recreation and cultural goods
1997	558	10	3
2005	731	9	7

Sources: ONS, *UK National Accounts*.

Normal, inferior and luxury goods

The income elasticity of demand measures how far the demand curve shifts horizontally when incomes change. Figure 4.5 shows two possible shifts caused by a given percentage increase in income. The income elasticity is larger if the given rise in income shifts the demand curve from DD to $D''D''$ than if the same income rise shifts the demand curve only from DD to $D'D'$. When an income rise shifts the demand curve to the left, the income elasticity of demand is a negative number, indicating that higher incomes are associated with smaller quantities demanded at any given prices.



Beginning at *A* on the demand curve *DD*, the income elasticity measures the horizontal shift in the demand curve when income rises 1 per cent. At the given price P_0 , a shift to *B* on the demand curve *D'D'* reflects a lower income elasticity than a shift to *C* on the demand curve *D''D''*. Leftward shifts in the demand curve when income rises indicate a negative income elasticity.

Figure 4.5 Income elasticity and shifts in demand

In Chapter 3 we distinguished **normal goods**, for which demand increases as income rises, and **inferior goods**, for which demand falls as income rises. We also distinguish between **luxury goods** and **necessities**. All inferior goods are necessities, since their income elasticities of demand are negative. However, necessities also include normal goods whose income elasticity of demand lies between zero and 1.

These definitions tell us what happens to budget shares when incomes are changed but prices remain unaltered. The budget share of inferior goods falls as incomes rise. Higher incomes and household budgets are associated with lower quantities demanded at constant prices. Conversely, the budget share of luxuries rises when income rises. Because the income elasticity of demand for luxuries exceeds 1, a 1 per cent rise in income increases quantity demanded (and hence total spending on luxury goods) by more than 1 percent. Rises in income reduce the budget share of normal goods that are necessities. A 1 percent income rise leads to a rise in quantity demanded but of less than 1 percent, so the budget share must fall.

A **normal good** has a positive income elasticity of demand.

An **inferior good** has a negative income elasticity of demand.

A **luxury good** has an income elasticity above unity.

A **necessity** has an income elasticity below unity.

Inferior goods tend to be goods for which there exist more expensive substitutes. Poor people satisfy their needs for meat and clothing by buying fatty meat and polyester shirts. As their incomes rise, they switch to better cuts of meat (steak) and more comfortable shirts (cotton). Rising incomes lead to an absolute decline in the demand for fatty meat and polyester shirts.

Luxury goods tend to be high-quality goods for which there exist lower-quality, barely adequate, substitutes: BMWs rather than small Fords, foreign rather than domestic holidays. Necessities that are normal goods lie between these two extremes. As incomes rise, the quantity of food demanded will rise but only a little. Most people still enjoy fairly simple home cooking even when their incomes rise. Looking back at Table 4.7, recreation and cultural goods are luxuries whose budget share increased from 3 to 7 per cent as UK incomes rose during 1997–2005. Food and drink cannot be a luxury, since its budget share fell as incomes rose, but it is not an inferior good either. At constant prices which adjust for the effects of inflation, during 1997–2007 real food spending *increased* from £56 billion (10 per cent of £558 billion) to £66 billion (9 per cent of £731 billion).

Table 4.9 summarizes the demand responses to changes in income holding constant the prices of all goods. The table shows the effect of income increases. Reductions in income have the opposite effect on quantity demanded and budget share.

Table 4.9 Demand responses to a 1% rise in income

Good	Income elasticity	Quantity demanded	Budget share	Example
Normal	Positive	Rises		
Luxury	Above 1	Rises more than 1%	Rises	BMW
Necessity	Between 0 and 1	Rises less than 1%	Falls	Food
Inferior	Negative	Falls	Falls	Bread

Table 4.10 reports estimated income elasticities of demand in the UK, for broad categories of goods in the first two columns and narrower categories in the last two columns. Again, the variation in elasticities is larger for narrower definitions of goods. According to those estimates, food is a normal good but not a luxury. Its income elasticity is 0.5.

The last column indicates that, within the food budget, higher income leads to a switch towards vegetables (whose income elasticity is higher than that for food as a whole) and away from bread. Rich households eat healthier food than low-income households. Notice that margarine and liquid wholemilk are inferior goods according to the results in Table 4.10.

Table 4.10 UK income elasticities of demand

Broad categories	Income elasticity	Narrower categories	Income elasticity
------------------	-------------------	---------------------	-------------------

Tobacco	0.5	Coal	2.0
Fuel	0.3	Bread and cereals	0.1
Food	0.5	Margarine	-0.37
Alcohol	1.1	Liquid wholemilk	-0.17
Clothing	1.2	Vegetables	0.9
Durables	1.5	Leisure goods	2.0
Services	1.8	Wines and spirits	2.6

Sources: J. Muellbauer, 'Testing the Barten model of household composition effects's, Economic Journal 87 (1977): 460–487; A. Deaton, 'The measurement of income and price elasticities's, European Economic Review 7 (1975): 261–273; National Food Survey 2000.

CASE 4.3

CAR CRAZY

As countries develop and get richer, one of the first things people want is a car. The income elasticity of demand for cars has been estimated at around 2. China, one of the fastest-growing economies in the world, now has an insatiable appetite for cars. With rapidly rising incomes, the once common bicycle is fast giving way to the car. In 1949 the world's most populous economy had a mere 1800 cars. By 2005 that figure was 24 million, making China second only to the US in the size of its car market. In 2009 China became the biggest market in the world in terms of cars. In that year, the number of cars sold in China increased by 49 per cent compared to 2008. In 2012 the number of new registered cars in China was 13.6 million, while in all Europe it was 12.5 million. Car ownership in China reached 114 million in 2012.

Whereas in the 1970s a worker had to save for a year to buy a bicycle, incomes are now so high that workers only have to save for a year to buy a car. Income elasticities of demand help us make confident predictions that rapidly rising living standards in countries such as China and India will lead to massive increases in the demand for cars, mobile phones, energy, air travel, and many other goods and services enjoyed in the affluent West. Since many of these are the source of emissions that lead to global warming, the very success of emerging economies adds new urgency to the need to find ways to reduce emissions, either by finding new, cleaner

technologies or by co-ordinated government policies to discourage the activities with which harmful emissions are associated.

Source: adapted from <http://news.bbc.co.uk/1/hi/business/6364195.stm> © bbc.co.uk/news.

Using income elasticities of demand

Income elasticities help us forecast the pattern of consumer demand as the economy grows and people get richer. Suppose real incomes grow by 15 per cent over the next five years. The estimates of Table 4.10 imply that margarine demand will fall by $5.55 (= 15 \times (-0.37))$ per cent, while the demand for wines and spirits will rise by 39 per cent. The growth prospects of these two industries are very different. These forecasts will affect decisions by firms about whether to build new factories and government projections of tax revenue from margarine and alcohol.

4.7 Inflation and demand

Elasticities measure the response of quantity demanded to separate variations in three factors: the own-price, the price of related goods and income. Chapter 2 distinguished *nominal* variables, measured in the prices of the day, and *real* variables, which adjust for inflation when comparing measurements at different dates. We end this chapter by examining the effect of inflation on demand behaviour.

Suppose all nominal variables double. Every good costs twice as much, wage rates are twice as high, rents charged by landlords and dividends paid by firms double in money terms. Whatever bundle of goods was previously affordable is still affordable. Goods cost twice as much but incomes are twice as high. If meat costs twice as much as bread, it still costs twice as much. Nothing has really changed. Demand behaviour will be unaltered by a doubling of the nominal value of *all* prices and *all* forms of income.

How do we reconcile this with the idea that own-price elasticities measure changes in quantity demanded as prices change? Each of the elasticities (own-price, cross-price and income) measures the effect of changing that variable *holding constant all other determinants of demand*. When all prices and all incomes are simultaneously changing, the definitions of elasticities warn us that it is incorrect to examine the

effect of one variable, such as the own-price, on quantity demanded. We can decompose the change in quantity demanded into three components: the effect of changes in the own-price alone, plus the effect of changes in price of other goods alone, plus the effect of changing incomes. When all nominal variables change by the same proportion, the sum of these three effects is exactly zero.

4.8 Elasticity of supply

Whereas the analysis of demand elasticities is quite tricky, the analysis of supply elasticities is refreshingly simple. We really need only keep track of the supply response to an increase in the own-price of a good or service. The elasticity of supply measures the responsiveness of the quantity supplied to a change in the price of that commodity.

Supply elasticity = (% change in quantity supplied)/(% change in price)

Because supply curves slope upwards, the elasticity of supply is *always positive*. As we move along a supply curve, positive price changes are associated with positive output changes. The more elastic is supply, the larger the percentage increase in quantity supplied in response to a given percentage change in price. Thus, elastic supply curves are relatively flat and inelastic supply curves relatively steep.

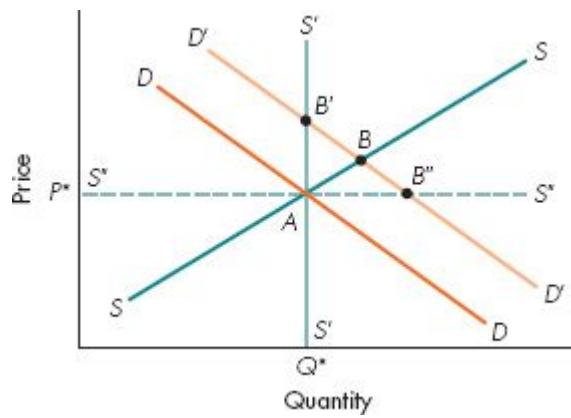


Figure 4.6 Alternative supply elasticities

Figure 4.6 shows a typical supply curve SS with a positive supply elasticity. If the supply curve is a straight line, the supply elasticity will change as we move along it. As we learned in relation to demand curves, a constant slope implies equal absolute changes in quantity as we successively increase price by one unit; however, these equal absolute changes imply different percentage changes, depending on the point from which we begin.

Figure 4.6 also shows two extreme cases. The vertical supply curve $S' S'$ has a zero supply elasticity. A given percentage change in price is associated with a zero percentage change in quantity supplied. The horizontal supply curve $S''S''$ has an infinite supply elasticity. Any price increase above the price P^* leads to an infinite increase in quantity supplied.

MATHS 4.3

THE POINT ELASTICITY OF SUPPLY

Just as we can for the elasticity of demand, we can also calculate the elasticity of supply at a given point of a specific supply function.

The price elasticity of supply (PES) is defined as:

$$PES = \frac{\% \Delta Q^S}{\% \Delta P} \quad (1)$$

The formula in (1) will measure the elasticity of supply between two different points of a given supply function. The point elasticity of supply is defined as:

$$PES = \frac{dQ^S}{dP} \frac{P}{Q^S} \quad (2)$$

where the term dQ^S/dP measures the slope of the supply function at a given point.

Consider the following linear direct supply function:

$$Q^S = 2 + 5P \quad (3)$$

where Q^S denotes the quantity supplied and P the price.

Suppose we want to find the elasticity of supply when $P = 10$. According to (3), at that price $Q^S = 52$. In this case, the slope of the supply function is constant and equal to 5. The point elasticity of supply when $P = 10$ and $Q^S = 52$ is therefore:

$$PES = 5 \times \frac{10}{10} = 0.96$$

The point elasticity of demand is less than 1. This means that a 1 per cent increase in the price will increase less than proportionally (by 0.96 per cent) the quantity supplied.

The elasticity of supply tells us how the equilibrium price and quantity will change when there is a shift in demand. Figure 4.6 shows a demand shift from DD to $D'D'$. Beginning from equilibrium at A , a demand shift from DD to $D'D'$ leads to a new equilibrium at B' , B or B'' depending on the elasticity of supply. The more inelastic is supply, the more the demand increase leads to higher prices rather than higher quantities. In the extreme case, the move from A to B' reflects only a price increase and the move from A to B'' reflects only a quantity increase. Table 4.11 provides a summary.

Table 4.11 Elasticities: a summary

	% change in quantity demanded	% change in quantity supplied
	induced by	
(Own-)Price elasticity of demand	1% rise in own price	
Cross-price elasticity of demand	1% rise in price of related good	
Income elasticity of demand	1% rise in income	
Elasticity of supply		1% rise in own price

Table 4.12 Elasticity of supply

Type of machinery	Elasticity of supply
Construction machinery	9.8
Farm machinery	2.7

Source: Edgerton J. (2010), ‘Estimating Machinery Supply Elasticities Using Output Price Booms’, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1750268.

Table 4.12 presents some estimates for the price elasticity of supply for machinery in the US. In all cases, the elasticity of supply is positive and larger than 1, meaning that the supply of that machinery appears to be elastic. For example, an increase of 1 per cent in the price of construction machinery will lead to an increase in the amount supplied of that machinery of almost 10 per cent.

4.9 Taxation and elasticity: who really pays the tax?

By spending and taxing, the government affects resource allocation in the economy. By taxing cigarettes, the government can reduce the number of cigarettes smoked and thereby improve health.

By taxing fuel, it can discourage pollution, though it may incur the wrath of lorry drivers and motorists. By taxing income earned from work, the government affects the amount of time people want to work. Taxes loom large in the workings of a mixed economy and have a profound effect on the way society allocates its scarce resources.

Initially we discuss what are called *specific* taxes, those that specify a particular amount, such as £5 per bottle of vodka. We show how the effect of a specific tax is related to the slope of supply and demand curves. We then extend the argument to *ad valorem* taxes, which are measured as a percentage of the commodity’s value. For example, VAT is usually levied at 20 per cent of the value of the good or service.

Just as specific taxes, in particular units, are related to slopes of supply and demand curves in particular units, so *ad valorem* or percentage taxes are related to *elasticities* of supply and demand, which are already expressed in percentages.

Either way, what we want to know is who ends up paying the tax. Suppose for simplicity we imagine a packet of cigarettes costs £1 and the government imposes a specific tax of 50p per packet. Do smokers end up paying the tax, or is it borne by cigarette producers? How much of the

tax can producers pass on to the consumer? We now show that this depends on the slopes of the supply and demand curves.

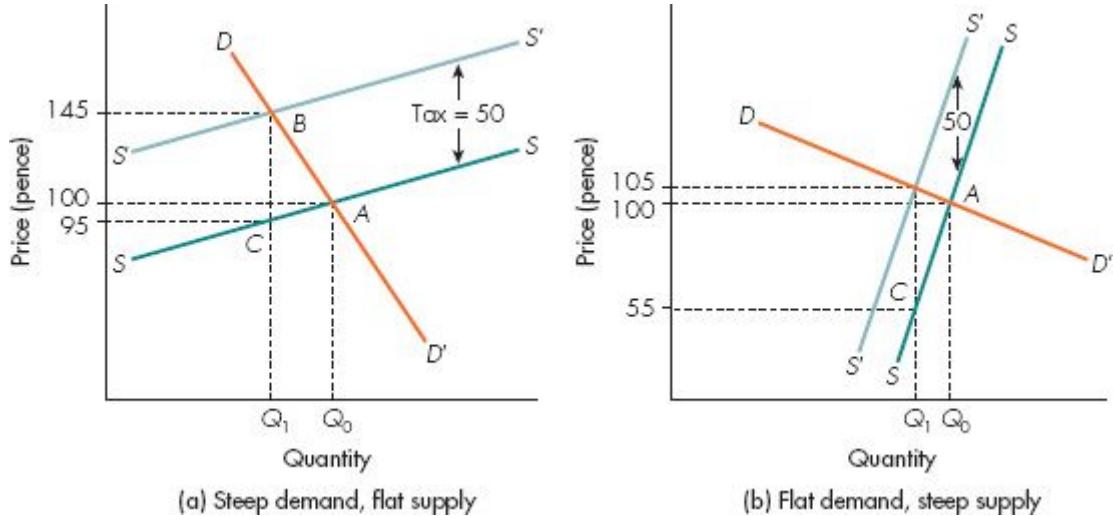


Figure 4.7 Tax incidence

Figures 4.7(a) and 4.7(b) plot the (after-tax) price to the consumer on the vertical axis. DD' shows the demand curve, which depends on the price to smokers (consumers). Since the price received by the producer is the consumer price minus the 50p tax per packet, the effect of the tax is to shift the supply curve from SS to S'S' in both diagrams. Each possible quantity supplied depends on the price received by the producer, which will be the same as before only if consumer prices are 50p higher: that is why we must shift the supply curve up by 50p.

In Figure 4.7(a), with a flat supply curve and steep demand curve, the tax is borne mainly by cigarette consumers. Point B is nearly 50p higher than point A. Since demand is insensitive to price, producers can pass on most of the tax in higher prices. Supply is price-sensitive, so the price received by producers cannot fall much. Consumers pay £1.45 and producers get £0.95 a packet.

In Figure 4.7(b), with a flat demand curve and a steep supply curve, most of the tax is borne by cigarette producers. Demand is price-sensitive, so attempts to pass on the tax in higher prices quickly lead to a drop in sales. Supply is price-insensitive and producers hardly cut back even though the price they receive has fallen by nearly 50p. Consumers pay £1.05 and producers get £0.55 a packet.

The key implication is thus that the **incidence of a tax** – who eventually bears the burden – has nothing to do with who initially hands over money to the government. The existence of the tax changes behaviour. This has induced effects on equilibrium price and quantity. These induced effects may be large or small, depending on the slopes of supply and demand curves.

The **incidence of a tax** describes who eventually bears the burden of that tax.

Now that we understand the general principle, it is obvious that the same argument will carry over to the more commonly used *ad valorem* taxes. We simply need to remember to confront the change in the percentage tax rate with the relevant percentage measures of price responsiveness of supply and demand; namely, the (own-price) supply and demand elasticities.

Hence, when demand is inelastic but supply elastic, the case for percentages corresponding to the absolute change case in Figure 4.7(a), a rise in an *ad valorem* tax will largely be borne by buyers in the form of a higher price paid. Conversely, when demand is elastic, but supply inelastic, the analogue of Figure 4.7(b), a higher *ad valorem* tax will fall mainly on suppliers, in the form of a lower price received. Thus, supply and demand elasticities help us think about the incidence of the commonest taxes, such as income tax, VAT and the corporation tax paid by companies, all of which are *ad valorem*.

MATHS 4.4

THE EFFECTS OF A SPECIFIC TAX

Here, we derive mathematically the effect of introducing a specific tax in a competitive market. With a specific tax, the tax bill depends on the quantity sold of a good and not on its price. Suppose that the market demand is linear and given by $Q^D = a - bP$. The market supply is given by $Q^S = c + dP$. Without any government intervention, the market outcome is:

$$P^* = \frac{(a - c)}{(b + d)} \text{ and } Q^* = \frac{(bc + da)}{(b + d)}$$

Now suppose that the government introduces a specific tax in the market that has to be paid by the suppliers. This creates a wedge between the price the consumers pay and that which the suppliers receive. In particular, when the specific tax is charged to the suppliers we have:

$$P^S = P^D - t$$

where t is the tax rate, P^S is the price received by the suppliers and P^D is the price paid by the consumers. In order to see the effects of this new tax, we modify the market demand and supply in the following way:

$$Q^D = a - bP^D \quad \text{and} \quad Q^S = c + dP^S$$

Now we take into account the fact that the price that affects the demand (the price consumers pay) can be different from the price that affects supply (the price the suppliers receive). Applying the knowledge that $P^S = P^D - t$ to the market supply, we have: $Q^S = c + d(P^D - t)$.

Notice that the introduction of the tax has the effect of shifting the market supply to the left by an amount given by t .

The market equilibrium is always where demand is equal to supply: $Q^D = Q^S$.

This implies: $a - bP^D = c + d(P^D - t)$. Solving that equation for P^D , we get:

$$P^{*D} = \frac{(a - c)}{(b + d)} + \frac{dt}{b + d} \quad (1)$$

This is the price paid in equilibrium by the consumers after the specific tax has been introduced in the market. The price received by suppliers is therefore: $P^S = P^D - t \Rightarrow [(a - c) + dt]/(d + b) - t$. Simplifying that equation, we obtain:

$$P^{*S} = \frac{(a - c)}{d + b} - \frac{bt}{b + d} \quad (2)$$

To find the equilibrium quantity, substitute equation (1) into the demand function (or equation (2) into the supply function):

$$Q^{*D} = \frac{ad + cb}{b + d} - \frac{bdt}{b + d}$$

What are the effects of the specific tax introduction? We can evaluate those effects by comparing the market equilibrium without the tax with the one once the tax is introduced (a comparative statics exercise).

Notice that:

$$\begin{aligned} P^{*D} &= P^* + \frac{dt}{b + d} \\ P^{*S} &= P^* - \frac{bt}{b + d} \\ Q^{*D} &= Q^* - \frac{bdt}{b + d} \end{aligned}$$

where P^* and Q^* were the equilibrium values before the tax introduction.

Therefore the tax introduction has the following effects on equilibrium:

- (1) It increases the price paid by consumers.
- (2) It decreases the price received by the suppliers.
- (3) It reduces the equilibrium quantity in the market.

Notice an important thing: it does not matter if the government charges the suppliers or the consumers, the results will be the same: $P^S = P^D - t$ implies that $P^D = P^S + t$. Using one equation instead of the other in our model will not change the final result.

Government revenues from the tax are given by the tax rate multiplied by the equilibrium quantity after the introduction of the tax, that is: tQ^{*D} .

Summary

- Unless otherwise specified, **the elasticity of demand** refers to the **own-price elasticity**. It measures the sensitivity of quantity demanded

to changes in the own-price of a good, holding constant the prices of other goods and income. Demand elasticities are negative since demand curves slope down. In general, the demand elasticity changes as we move along a given demand curve. Along a straight-line demand curve, elasticity falls as price falls.

- **Demand is elastic** if the price elasticity is more negative than -1 (for example, -2). Price cuts then increase total spending on the good. **Demand is inelastic** if the demand elasticity lies between -1 and 0 . Price cuts then reduce total spending on the good. **Demand is unit-elastic** if the demand elasticity is -1 . Price changes then have no effect on total spending on the good.
- The demand elasticity depends on how long customers have to adjust to a price change. In the short run, substitution possibilities may be limited. Demand elasticities will typically rise (become more negative) with the length of time allowed for adjustment. The time required for complete adjustment varies from good to good.
- The **cross-price elasticity of demand** measures the sensitivity of quantity demanded of one good to changes in the price of a related good. Positive cross-price elasticities tend to imply that goods are **substitutes**, negative cross-price elasticities that goods are **complements**.
- The **income elasticity of demand** measures the sensitivity of quantity demanded to changes in income, holding constant the prices of all goods.
- **Inferior goods** have negative income elasticities of demand. Higher incomes reduce the quantity demanded and the budget share of such goods. **Luxury goods** have income elasticities larger than 1 . Higher incomes raise the quantity demanded and the budget share of such goods.
- Goods that are not inferior are called **normal goods** and have positive income elasticities of demand. Goods that are not luxuries are called

necessities and have income elasticities of less than 1. All inferior goods are necessities but normal goods are necessities only if they are not luxuries.

- Doubling all nominal variables should have no effect on demand since it alters neither the real value (purchasing power) of incomes nor the relative prices of goods. In examining data from economies experiencing inflation, it is often best to look at real prices and real incomes, adjusting prices and incomes for the effect of inflation.
- The **supply elasticity** measures the percentage response of quantity supplied to a 1 per cent increase in the price of the commodity. Since supply curves slope up, the supply elasticity is positive.
- **Tax incidence** measures who eventually pays the tax. Since taxes induce changes in equilibrium prices and quantities, this can be very different from the people from whom the government appears to collect the money.
- For **specific taxes**, slopes of supply and demand curves are relevant. For **ad valorem taxes**, elasticities of supply and demand are relevant. In either case, it is the more price-insensitive side of the market that bears more of the burden of a tax.

Review questions



EASY

- 1 Your fruit stall has 100 ripe peaches that must be sold at once. Your supply curve for peaches is vertical. From past experience, 100 peaches are demanded if the price is £1. The demand elasticity is -0.5 . (a) Draw a supply and demand diagram showing market equilibrium. (b) You discover 10 of your peaches are rotten and cannot be sold. Draw the new supply curve. What is the new equilibrium price?
- 2 For each of categories (a), (b) and (c), do you expect demand to be elastic or inelastic? (a) Milk, dental services, beer; (b) chocolate, chicken, train journeys; (c) theatre trips, tennis clubs, films. Then rank the elasticities within each category. Explain your answers.

- 3 Where along a straight-line demand curve does consumer spending reach a maximum? Explain why. What use is this information to the owner of a football club?
- 4 The following table shows price and income elasticities for vegetables and catering services. For each good, explain whether it is a luxury or a necessity, and whether demand is elastic or inelastic.

	Price elasticity	Income elasticity
Vegetables	0.17	0.87
Catering services	2.61	1.64

- 5 Common fallacies Why are these statements wrong? (a) Because cigarettes are a necessity, tax revenues from cigarettes will always increase when the tax rate is raised. (b) Farmers should take out insurance against bad weather that might destroy half of all their crops. (c) Higher consumer incomes always benefit producers.

MEDIUM

- 6 The data below refer to the market for cheese:

Quantity	Price
130	10
110	20
80	35
70	40
58	46
50	50

Plot the demand for cheese. Given that the demand for cheese is unit elastic at 37p, for which prices is the demand for cheese elastic? For which ones is the demand for cheese inelastic?

- 7 Consider the following demand function: $Q^D = 25/P^2$. Show that the point elasticity of demand for that function is always equal to -2 .
- 8 The data below refers to the quantity demanded of good A and the price of A as a result of changes to the price of good B and good C:

Q_A (kg)	P_A (pence)	P_B (pence)	P_C (pence)

3	52	32	64
1.3	82	26	71

Are goods A and B substitutes or complements? What about goods A and C?

- 9 Air conditioners are a luxury good. (a) What does this imply about income elasticity? (b) Which two countries would you guess have the highest per capita demand for air conditioners at present? (c) If people continue to get richer and global warming continues to increase, what is likely to happen to the quantity of air conditioners demanded? And what will this do to global warming? And hence to the demand for air conditioners? (d) Could this process spiral out of control?
- 10 Suppose the government levies a tax of 55p per bottle of wine. Show on a graph the effect of this tax on the equilibrium price and quantity assuming that the demand for wine is elastic. Do consumers or suppliers bear a greater proportion of the tax?
- 11 (a) If the government wants to maximize revenue from cigarette tax, should it simply set a very high tax rate on cigarettes? (b) If the government achieves its objective, what is the elasticity of demand for cigarettes at the price corresponding to this tax rate? You may assume that cigarettes are essentially free to produce and the entire price reflects the tax. (c) A research company measures elasticity and concludes that the demand for cigarettes is price-elastic. Should the government raise or lower the tax rate? (d) If the government wants to get some tax revenue but also wants to make people smoke less, should it set a tax rate above or below that which maximizes revenue from cigarette taxation?

MEDIUM

- 12 **Essay question** Suppose climate change causes flooding that wipes out much of UK agriculture. Discuss what happens to the price of food in the UK: (a) in the short run and (b) in the long run. Did you assume that the UK made and consumed all food itself or did you allow for international trade? How does the outcome differ in these two cases?

HARD

- |3 Suppose that the market demand for beef is given by $Q^D = 200 - 6P + 2Y$, where P is the price of meat per kg and Y is the consumers' income. Suppose that consumers' income is £100. If the price of beef decreases from £10 to £8 per kg, find the corresponding elasticity of demand. Now suppose that the price is fixed to £8 while consumers' income increases from £100 to £150; find the corresponding income elasticity of demand. Is beef a normal good?
- |4 The market demand for a given good is $Q^D = 26 - 4P$, while the market supply is $Q^S = 2P - 4$. Find the equilibrium price and quantity in the market. Now assume that the government introduces a specific tax $t = 3$ on the suppliers. Find the new equilibrium price and the new equilibrium quantity. Compare the pre-tax equilibrium with the after-tax equilibrium. What are the main differences?

- 1 From Figure 4.1 we can see that the slope of the demand curve is 20.125. Therefore $1/(20.125) = -8$.
- 2 Notice the difference between the price elasticity of demand and the slope of the demand. The slope of a demand curve is the ratio of the change in price to the change in quantity between two points on the curve. The price elasticity of demand is the ratio of the percentage change in quantity to the percentage change in price. You will see, for a linear demand curve, the elasticity changes along the demand curve, even if the slope remains constant. See Maths 4.2.
- 3 Except in two special cases: a horizontal, or infinitely elastic, demand curve has an elasticity of minus infinity at all points since the price change is always zero. A vertical, or completely inelastic, demand curve has an elasticity of zero at all points since the quantity never changes.
- 4 www.economist.co.uk, 30 July 1994.
- 5 IATA, Estimating air travel demand elasticities: Final report, 2007.
- 6 The concepts of substitutes and complements are discussed in more detail in Section 5.5.

CHAPTER 5

Consumer choice and demand decisions

Learning outcomes

By the end of this chapter, you should be able to:

- 1 define the relationship between utility and tastes for a consumer
- 2 describe the concept of diminishing marginal utility
- 3 describe the concept of diminishing marginal rate of substitution
- 4 understand how to represent tastes as indifference curves
- 5 understand how to derive a budget line
- 6 realize how indifference curves and budget lines explain consumer choice
- 7 describe how consumer income affects quantity demanded
- 8 understand how a price change affects quantity demanded
- 9 define income and substitution effects
- 10 recognize the market demand curve

In previous chapters we introduced demand curves to represent consumer behaviour. In this chapter we will build a formal theory of consumer choice to explain where those demand curves come from. This theory will help us to explain how consumers reconcile what they would like to consume, as described by their tastes or preferences, with what the market will allow them to do, as described by their incomes and the prices of various goods. We will then use this theory to predict how consumers will respond to changes in market conditions and to derive a consumer's demand curve. Furthermore, we will relate the theory developed in this chapter to the price and income elasticities examined in Chapter 4.

5.1 Demand by a single consumer

In order to build a theory of consumer choice, first we need to specify what the main objective of such a theory is: we want to explain how a consumer chooses how much to consume of different goods given her resources (income) and given the market conditions (the prices of the different goods). If we are able to identify the basic mechanism behind the decision process for a consumer, then we are able to understand how the quantity of goods bought (and so consumed) by that consumer can change if prices change. But this relationship between the quantity bought of a good and its price is what we call a demand function. Therefore our theory of consumer choice will provide a foundation for the demand curves we have studied in previous chapters.

The main ingredients of this theory are:

- *The consumer's tastes and utility* Tastes, or preferences, are the driving force behind what a consumer chooses to consume. Some prefer Coca-Cola to orange juice, some prefer beef to chicken, and so on. Utility is what economists call the satisfaction consumers get from consuming goods.
- *The behavioural assumption that consumers are rational* By rational, we mean that consumers will try to obtain the best they can from their consumption decisions. In particular, of the affordable consumption bundles, a rational consumer picks the bundle that maximizes her own satisfaction.
- *The consumer's income* This represents the resource available to the consumer for the consumption activity. A consumer cannot consume more than her available income.
- *The prices at which goods can be bought.*

Each element listed above is explained in detail in the following sections.

Tastes and utility

Consumers have tastes, or preferences, about the goods they consume. In building our theory of consumer choice we need to make some assumptions about those tastes. Before doing that we introduce the idea of a *consumption bundle*. This represents what a consumer would like to consume. A consumption bundle contains different quantities of various goods. In the following we make the simplifying assumption that a consumption bundle contains different quantities of only two goods: films and meals. For example, suppose that a consumer faces the following three bundles: bundle *a* contains 6 films and 5 meals; bundle *b* contains 4 films and 2 meals; and bundle *c* contains 2 films and 1 meal. The problem for our consumer is how to choose between those bundles. If she likes both films and meals a lot, she would probably go for bundle *a*, which contains more of both goods compared to the other bundles. However, without making any assumptions regarding how the consumer can rank different consumption bundles according to her tastes, we cannot say much. Before turning to the problem of taste, however, we introduce the concept of *utility*.

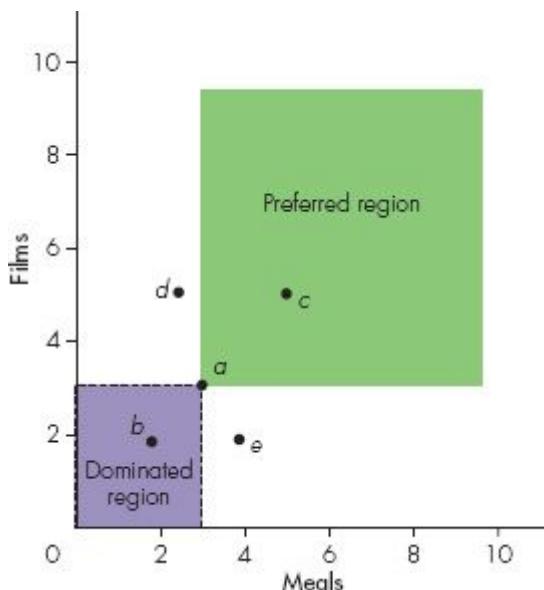
Consumers obtain utility from consuming goods. Therefore utility is the final objective of consumption. It represents what a consumer achieves by consuming a particular consumption bundle. For example, suppose you much prefer Italian food to French food. Then you would probably be *happier* going to eat in an Italian restaurant than in a French restaurant. Or to put it another way: you would probably get more *utility* from consuming the Italian food than the French food. From this simple example you can probably see that there is a link between the concept of utility and the tastes of a consumer. A consumer prefers one bundle of goods to another if the utility she gets from the former is greater than the utility she gets from the latter.

Given the concepts of consumption bundles and utility just defined, we now make the following plausible assumptions regarding the tastes of a consumer:

- *Completeness* The consumer can always rank alternative bundles of goods according to the satisfaction or utility they provide. It is unnecessary to quantify this utility,¹ for example to say that one bundle yields twice as much utility as another bundle. We only require that the consumer can decide that one bundle is better than, worse than or exactly as good as another. This assumption rules out

the possibility that a consumer facing different bundles cannot decide which one she prefers.

- *Transitivity* We assume that the ranking of possible bundles is internally consistent: if bundle a is preferred to bundle b and bundle b is preferred to bundle c , then bundle a must be preferred to bundle c .
- *Consumers prefer more to less* If bundle b offers more films but as many meals as bundle c , we assume bundle b is preferred. The same applies if bundle b offers more meals but as many films as bundle c . What about things, such as pollution, which are not goods but ‘bads’? Consumers do not prefer more pollution to less. We get round this problem by redefining commodities so that our assumption is satisfied. We analyse clean water rather than polluted water. More clean water is better than less.



The consumer evaluates consumption bundles a, b, c, d and e. With respect to point a, any point to the north-east is preferred and any point to the south-west is dominated by c. Points such as d or e in the other two regions may or may not be preferred to a, depending on the consumer’s tastes.

Figure 5.1 Ranking alternative consumption bundles

Figure 5.1 examines the implications of these three assumptions about taste. Each point shows a consumption bundle of films and meals. Now it is clear why we assumed that a consumption bundle contains only two goods. This assumption allows us to represent our analysis using nice and simple graphs, since a consumption bundle can be depicted as a point where the co-ordinates are the quantities of the two goods. For example, in Figure 5.1 point a represents a bundle containing 3 films and 3 meals.² We begin at bundle a . Since more is preferred to less, any point such as c to the north-east of a is preferred to a . Point c offers more of *both* goods than a . Conversely, points to the south-west of a offer less of both goods than a . Point a is preferred to points such as b . By transitivity, since bundle c is preferred to bundle a and bundle a is preferred to bundle b , then it must be that bundle c is preferred to bundle b . Notice that, given the assumptions made about the tastes of our consumer, we cannot be sure how points in the other two regions (north-west, south-east) compare with a . At d or e , the consumer has more of one good but less of the other good than at a . Someone who really likes food might prefer e to a , but an avid film buff would prefer d to a .

The **marginal rate of substitution (MRS)** between two goods measures the quantity of a good the consumer must sacrifice to increase the quantity of the other good by one unit *without changing total utility*.

Consumers prefer more to less. An extra meal increases utility. To hold utility constant when a meal is added, the consumer must sacrifice some of the other good (films). The **marginal rate of substitution (MRS)** tells us how many films the consumer is willing to exchange for an additional meal without changing total utility.³

Suppose the student has 5 films and no meals. Having already seen 4 films, she does not enjoy the fifth film much. With no meals, she is *very* hungry. The utility of this bundle (5 films and zero meals) is low: being so hungry, she cannot enjoy films anyway. For the same low amount of utility she could give up a lot of films for a little food.

Suppose instead that she eats a lot of meals but sees few films. She is then reluctant to sacrifice much cinema attendance to gain yet another meal. Previously, it made sense to sacrifice abundant films for scarce meals. Now, when the ratio of films to meals is already low, it does not make sense to sacrifice scarce films for yet more meals.

Consumer tastes exhibit a **diminishing marginal rate of substitution** when, to hold utility constant, diminishing quantities of one good must be sacrificed to obtain successive equal increases in the quantity of the other good.

This commonsense reasoning about tastes or preferences is very robust. It can become a general principle about consumer tastes. In general, the tastes of a consumer will display a property called **diminishing marginal rate of substitution**. The idea of preferences displaying diminishing marginal rate of substitution captures the fact that, when a consumer has a lot of one good, she is willing to give up a relatively large amount of it to get a good of which she has relatively little.

Our consumer might be equally happy with bundle X (6 films, 0 meals), bundle Y (3 films, 1 meal) and bundle Z (2 films, 2 meals). Beginning from bundle X , a move to Y sacrifices 3 films for 1 meal, but a further move from Y to Z sacrifices only 1 film for 1 extra meal. Such tastes satisfy the assumption of a diminishing marginal rate of substitution.

The assumptions we have made so far regarding consumers' tastes are all we require. It is now convenient to show how tastes can be represented as *indifference curves*.

Representing tastes as indifference curves

Indifference curves are a graphical representation of the tastes of a consumer. An **indifference curve** is defined as the curve representing all the combinations of consumption bundles that provide the same level of utility for a consumer. Therefore a consumer is indifferent between consuming any of the bundles that lie on a given indifference curve.

An **indifference curve** shows all the consumption bundles yielding a particular level of utility.

Figure 5.2 illustrates how an indifference curve should look. Start at bundle a . Bundles in the preferred region of Figure 5.2 (like bundle c) are preferred to a and therefore they provide a higher utility to our consumer. Therefore those bundles cannot be on the same indifference curve as bundle a . Bundles that are on the dominated region (like bundle

b) are worse than bundle *a* and therefore they provide lower utility to our consumer. Those bundles cannot be on the same indifference curve as bundle *a*. The only possible bundles that can provide the same utility to our consumer as bundle *a* are bundles that lie on the north-west and south-east regions compared with *a*.

An indifference curve is therefore a downward-sloping curve connecting all the bundles that our consumer considers as equally desirable in terms of the utility they provide.

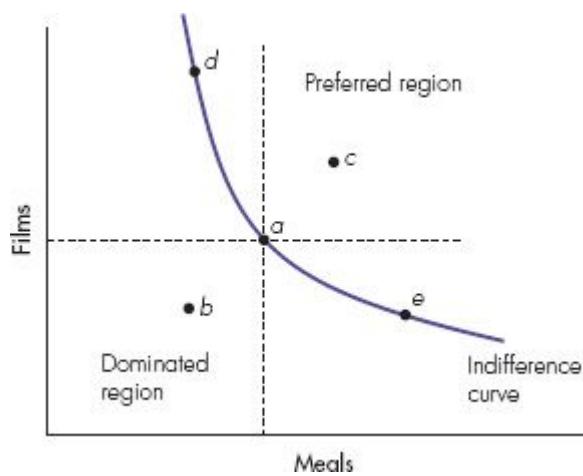
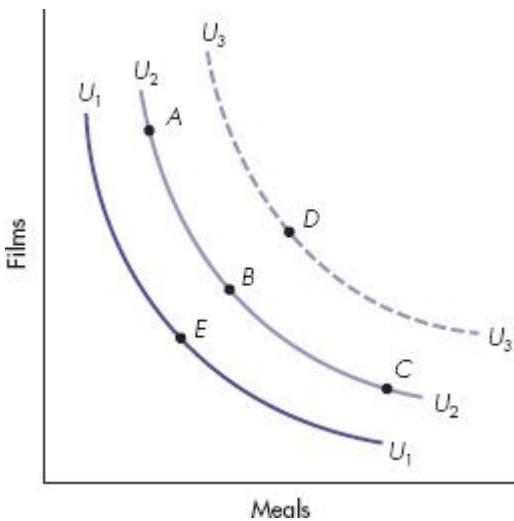


Figure 5.2 The typical shape of an indifference curve

Obviously, since the consumer can face many different bundles, we can have many different indifference curves on the same graph to represent the tastes of our consumer. Figure 5.3 shows three possible indifference curves, $U_1 U_1$, $U_2 U_2$ and $U_3 U_3$.



Along each curve consumer utility is constant. Since more is preferred to less, any point on a higher indifference curve is preferred to any point on a lower indifference curve. Indifference curves slope downwards. Otherwise the consumer would have more of both goods and be better off. Diminishing marginal rates of substitution imply that each curve becomes flatter as we move along it to the right.

Figure 5.3 Representing consumer tastes by indifference curves

Why do indifference curves represent graphically the tastes of a consumer? We show that the indifference curves displayed in Figure 5.3 satisfy all the assumptions we have made about consumers' tastes.

Consider the indifference curve denoted by U_2U_2 . By definition, every point on U_2U_2 yields the same utility for our consumer. Another way to say that is: our consumer is indifferent among all points on that indifference curve. Point C has many meals and few films, and point A offers many films but few meals. Because a consumer prefers more to less, *indifference curves must slope downwards*. Since more meals tend to increase utility, some films must simultaneously be sacrificed to hold utility constant.

The slope of a typical indifference curve gets steadily flatter as we move to the right. This reflects a diminishing marginal rate of substitution. At A, the consumer has relatively more films than meals. To get extra meals and moving, for example, to point B, she is willing to sacrifice a lot of films. At B, the quantities of films and meals are roughly the same. To get extra meals and move to point C, for example, she is willing to sacrifice fewer films. At C, she has so many meals that hardly any films

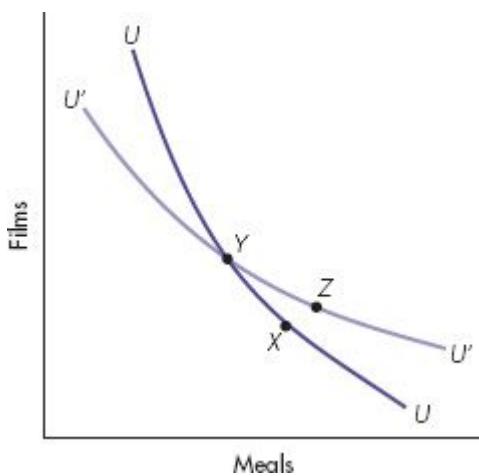
will be sacrificed for extra meals. The marginal rate of substitution of meals for films is simply the slope of the indifference curve at the point from which we began. These two properties of a single indifference curve – its downward slope and its steady flattening as we move to the right – follow directly from the assumption that consumers prefer more to less and from the fact that tastes generally display the property of diminishing marginal rates of substitution.

Now consider point D on indifference curve U_3U_3 . D offers more of both goods than B . Since consumers prefer more to less, utility at D is higher than utility at B . But all points on U_3U_3 yield the same utility as each other. Thus, every point on U_3U_3 yields more utility than every point on U_2U_2 . Conversely, E must yield less utility than B since it offers less of both goods. Every point on U_1U_1 yields less utility than every point on U_2U_2 .

Although Figure 5.3 shows only three indifference curves, we can draw in other indifference curves as well. In particular, there is an indifference curve passing through every possible bundle faced by our consumer. This comes from the assumption of completeness of tastes.

Indifference curves further away from the origin (higher indifference curves) are associated with higher levels of utility because the consumer prefers more to less.

Indifference curves cannot cross. Figure 5.4 shows why. Suppose UU and $U'U'$ cross. Since X and Y lie on the indifference curve UU , the consumer is indifferent between these points. But Y and Z lie on the indifference curve $U'U'$. Hence, the consumer is indifferent between Y and Z . Hence, the consumer is also indifferent between X and Z . This is impossible, since the consumer gets more of both goods at Z than at X . Intersecting indifference curves would violate our assumption that consumers prefer more to less.



If indifference curves intersected, the consumer would be indifferent between X and Y on the indifference curve UU' , and between Y and Z on $U'U$, and hence indifferent between X and Z. Since Z offers more of both goods than X, this violates the assumption that consumers prefer more to less. Indifference curves cannot intersect.

Figure 5.4 Indifference curves cannot intersect

Furthermore if indifference curves cross, the assumption regarding the transitivity of tastes will not hold. If X is indifferent to Y and Y is indifferent to Z , then by the logical consistency of tastes we should have that X is indifferent to Z . However, we know that Z should be preferred to X because of the assumption that consumers prefer more to less.

Our assumptions about consumer tastes rule out intersecting indifference curves.

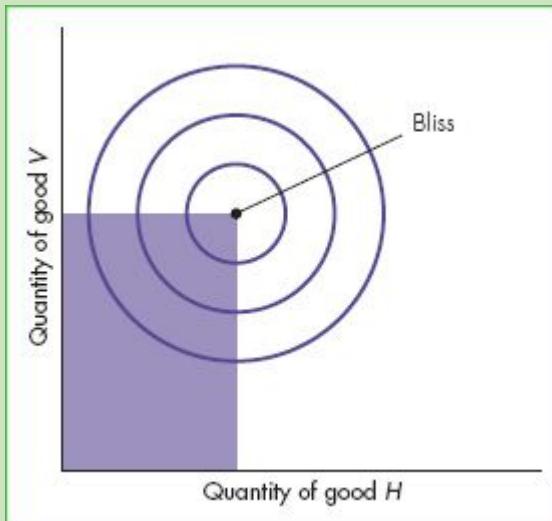
CONCEPT 5.1

OTHER CONTOUR MAPS

When you look at a good map, you will see concentric rings or contours, each showing points of equal height. They are like indifference curves but do not have to obey the law of diminishing marginal rate of substitution and hence have stranger shapes. But they never intersect. Different contours are different heights.

As you rise through successive contours, you reach a dot marking the top of the mountain. In economics, we hardly ever reach the top. People are rarely satiated. But an indifference map for

champagne and lobster might look like a mountain. Too much of either, and you are sick. The dot for the absolute best, or bliss, point, equivalent to the top of the mountain, then shows the finite combination of champagne and lobster preferred above all others, however much is available. We are now violating our assumption that the consumer always prefers more to less. The part of the contour sloping up corresponds to the range in which more is no longer better.



For broad categories of commodities, we are never satiated. It is as if we are confined to the shaded area of the diagram.

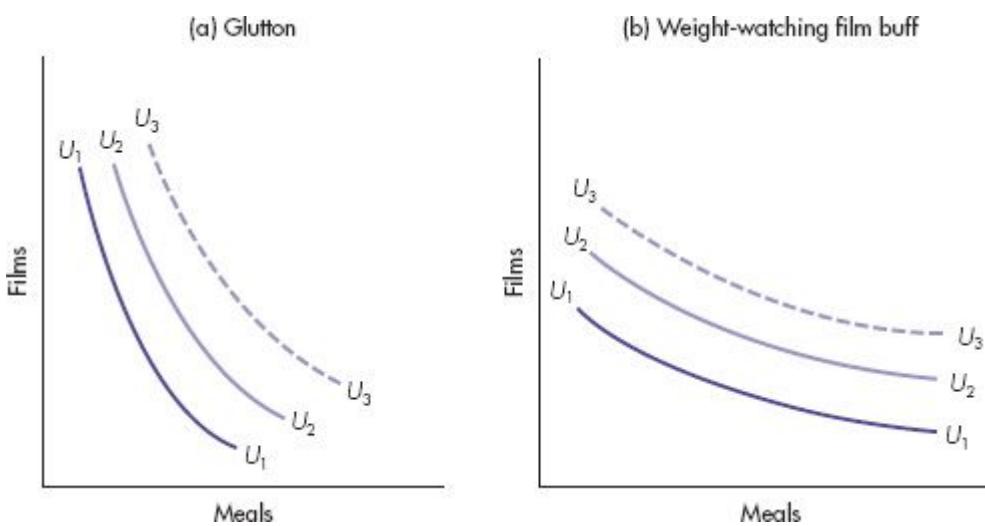
We can represent the tastes of any consumer by drawing the complete *map* of her indifference curves. Figure 5.5 shows two consumers with different tastes. In each case, moves to a higher indifference curve imply an increase in utility. Figure 5.5(a) shows the indifference map for a glutton prepared to give up a lot of films to gain a little extra food. Figure 5.5(b) shows the indifference map for a weight-watching film buff, who will give up large quantities of food to see more films. Both indifference maps are valid: they satisfy our basic assumptions about consumer tastes. Our theory can cope with extreme preferences as well as with more typical preferences in-between.

The budget constraint

A consumer's income and the market prices of goods define her **budget constraint**.⁴ The budget constraint introduces the problem of scarcity into our analysis.

Consider a student with a weekly budget of £50 to be spent on meals or films. Each meal costs £5 and each film £10. Those are the market prices of the two goods. We assume that the consumer takes those prices as given.⁵ What combination of meals and films can she afford? Going without films, she can spend £50 on 10 meals at £5 each. Going without meals, she can buy 5 cinema tickets at £10 each. Between these two extremes lie many combinations of meals and films that together cost exactly £50. These combinations are called the budget constraint.

The **budget constraint** describes the different bundles that the consumer can afford.



Holding utility constant on any indifference curve, to get more food a glutton sacrifices more films than a film buff.

Figure 5.5 Different tastes

The budget constraint shows the *maximum* affordable quantity of one good given the quantity of the other good being purchased.⁶ Table 5.1 shows her budget constraint. Each row shows a bundle whose total value of £50 just exhausts her income.

Table 5.1 Affordable consumption baskets

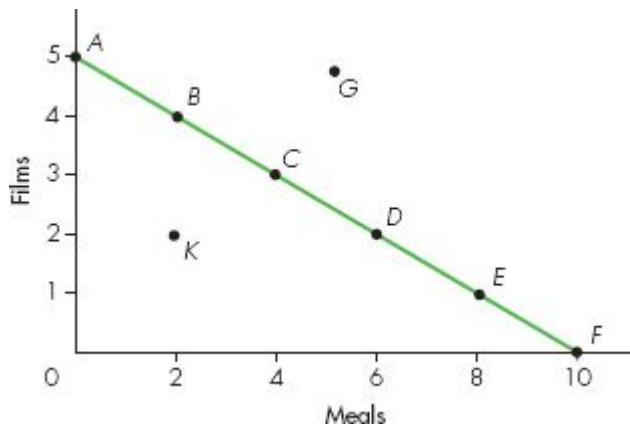
Quantity of meals Q_M	Spending on meals £5 $\times Q_M$	Quantity of films Q_F	Spending on films £10 $\times Q_F$	Total spending £
0	0	5	50	50
2	10	4	40	50
4	20	3	30	50
6	30	2	20	50
8	40	1	10	50
10	50	0	0	50

Table 5.1 shows the *trade-off* between meals and films. Higher quantities of meals require lower quantities of films. For a given income, the budget constraint shows how much of one good must be sacrificed to obtain larger quantities of the other good. It is because there is a trade-off that she must *choose between* meals and films.

When the price of meals and films is fixed, independently of how many she buys, her budget constraint is a straight line, sometimes called the budget line. Figure 5.6 plots this budget line using the budget constraint data of Table 5.1.

The position of the budget line is determined by its end-points A and F , which have a simple interpretation. Point A is the most films the budget will buy if the student has no meals: £50 buys at most 5 film tickets at £10 each. Point F shows that £50 buys at most 10 meals at £5 each if she has no films. The budget line joins up points A and F . Intermediate points such as B and C show more balanced purchases of meals and films.

The slope of the budget line shows how many meals must be sacrificed to get another film. Thus it represents the *opportunity cost* of films in terms of meals. Moving from point F to point E reduces the quantity of meals from 10 to 8 but raises the quantity of films from 0 to 1. This trade-off between meals and films is constant along this budget line. Giving up 2 meals always yields the extra £10 to buy 1 extra film.



The budget line shows the maximum combinations of goods that the consumer can afford, given income and the prevailing prices. Points on the budget line use up the entire consumer budget. Points above the budget line are unaffordable. Points inside the budget line would allow additional spending.

Figure 5.6 The budget line

Since films cost twice as much as meals, 2 meals must be sacrificed to buy 1 more film ticket. *The slope of the budget line depends only on the ratio of the prices of the two goods.* The slope of a line is the change in the vertical distance divided by the corresponding change in the horizontal distance. In Figure 5.5 the slope of the budget line is $-1/2$. The (+1) change in films is divided by the (-2) change in meals. This example illustrates the general rule

$$\text{Slope of the budget line} = -P_H/P_V$$

where P_H is the price of the good on the horizontal axis and P_V is the price of the good on the vertical axis. In our example, the price of meals is $P_H = £5$ and the price of films is $P_V = £10$. The formula confirms that the slope of the budget line is $-1/2$. The minus sign reminds us that there is a trade-off. We have to *give up* one good to get more of the other good.

The two end-points of the budget line (here, A and F) show how much of each good the budget buys if the other good is not bought at all. The slope of the budget line joining these end-points depends only on the relative prices of the two goods.

Any point above the budget line (such as G in Figure 5.6) is unaffordable. The budget line shows the maximum quantity of one good

that is affordable, given the quantity of the other good purchased and the budget available to spend. With an income of £50, G is out of reach: it would need £25 to buy 5 meals and £50 to buy 5 cinema tickets. Points such as K , which lie inside the budget line, leave some income unspent. Only on the budget line is there a trade-off where the student must choose *between* films and meals.

MATHS 5.1

THE BUDGET CONSTRAINT AND THE BUDGET LINE

It is helpful to see how we can express the budget constraint in a mathematical way. A consumer faces different bundles containing different quantities of two goods; call them X and Y . The consumer takes the prices of the two goods as given. Define by p_X the price of good X and by p_Y the price of good Y . Define by x the quantity of good X and by y the quantity of good Y . Define by M the income available to the consumer. Then the budget constraint of the consumer can be written as:

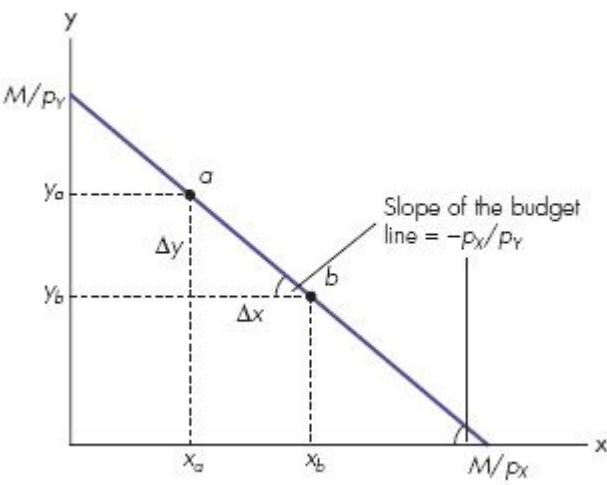
$$p_X x + p_Y y = M \quad (1)$$

The left-hand side of the equation above is the total expenditure of the consumer who buys quantities x and y of the two goods at the given prices. The expenditure of the consumer must be equal to her income M .

From the budget constraint in equation (+) we can derive the corresponding budget line using simple algebra:

$$y = \frac{M}{p_Y} - \frac{p_X}{p_Y} x \quad (2)$$

Equation (2) is the budget line. The slope of the budget line is given by $-p_X/p_Y$; that is, the price ratio.



Since the budget line is a straight line, the slope is constant along the line. The term M/p_Y is the vertical intercept of the budget line. It tells you how much the consumer can buy of good Y when she spends all her income on good Y .

The term M/p_X is the horizontal intercept of the budget line; it tells you the amount of good X that the consumer can buy if she spends all the income on good X . Bundle a on the graph contains an amount x_a of good X and an amount y_a of good Y . In moving from bundle a to bundle b , the consumer increases her consumption of good X , from x_a to x_b , but she has to reduce the consumption of good Y from y_a to y_b . This shows the tradeoff between the two goods implied by the budget constraint. Since the budget line is linear, we can find the slope of the budget line analytically in a simple way. Notice that both bundles (a and b) satisfy the budget constraint; therefore it must be true that:

$$y_a = \frac{M}{p_Y} - \frac{p_X}{p_Y} x_a \quad (3)$$

and

$$y_b = \frac{M}{p_Y} - \frac{p_X}{p_Y} x_b \quad (4)$$

By taking the difference between (4) and (3), we obtain:

$$y_b - y_a = \frac{M}{p_Y} - \frac{M}{p_Y} - \frac{p_X}{p_Y} x_b + \frac{p_X}{p_Y} x_a \quad (5)$$

Call $\Delta y = y_b - y_a$ the change in the quantity of Y between bundle b and bundle a , and similarly for good X define $\Delta x = x_b - x_a$. Then (5) can be written as:

$$\Delta y = -\frac{p_X}{p_Y} \Delta x$$

Or, written differently:

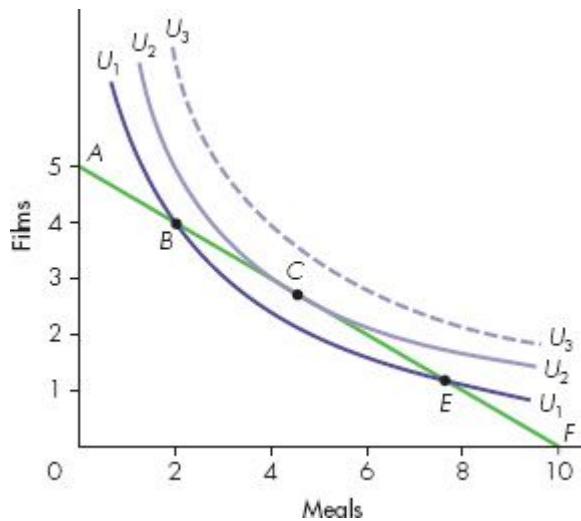
$$\frac{\Delta y}{\Delta x} = -\frac{p_X}{p_Y}$$

that is, exactly the slope of the budget line.

Utility maximization and choice

The budget line shows affordable bundles given a consumer's market environment (her budget and the price of different goods). The indifference map shows her tastes. We put the two concepts together in order to see how the consumer chooses among different bundles. We assume that the consumer is *rational*, implying that she *chooses the affordable bundle that maximizes her utility*. She cannot afford points above the budget line and will never choose points below the budget line (it is then possible to buy more of one good without sacrificing any of the other good). She will select a point on the budget line – her **chosen bundle**.

The **chosen bundle** will be the point at which an indifference curve just touches the budget line. The budget line is a tangent to the indifference curve at this point.



Points above the budget line AF are unaffordable. The consumer cannot reach the indifference curve U_3U_3 . Points such as B and E are affordable but only allow the consumer to reach the indifference curve U_1U_1 . The consumer will choose the point C to reach the highest possible indifference curve U_2U_2 . At point C , the indifference curve and the budget line just touch and their slopes are equal.

Figure 5.7 Consumer choice in action

To find which point on the budget line maximizes utility, we examine the consumer's tastes. Our glutton should pick a point with more meals and fewer films than the point our film buff selects. We first show how to use indifference curves to find the bundle the consumer chooses. Then we confirm that our model of consumer choice captures the different behaviour of the glutton and the film buff.

Figure 5.7 shows the budget line AF for the student who had £50 to spend on films (£10 each) and meals (£5 each). The indifference curves U_1U_1 , U_2U_2 and U_3U_3 are part of the indifference map describing her tastes.

All points on U_3U_3 are unattainable since it lies entirely above the budget line AF . The student would like this high level of utility but cannot afford it. Next, suppose she considers the attainable point B on the indifference curve U_1U_1 . She prefers this to point A , which must lie on a lower indifference curve (since indifference curves cannot intersect, the indifference curve through A lies entirely below the indifference

curve U_1U_1). Similarly, F must lie on a lower indifference curve than E and she prefers E to F .

However, she will choose neither B nor E . By moving to C , she reaches a higher indifference curve and gets more utility, and we have assumed that the consumer chooses in order to obtain the highest utility. C is the point she chooses. Any other affordable point on the budget line is on a lower indifference curve. The budget line never crosses a higher indifference curve, such as U_3U_3 , and crosses twice every lower indifference curve, such as U_1U_1 . Point C is the point of maximum utility given the budget constraint.

We can reach the same answer by different means. Consider again point B in Figure 5.7. The slope of the budget line shows the trade-off between affordable quantities of films and meals that the market environment will allow. When films cost £10 and meals £5, two meals can be traded for one film. The slope of the indifference curve at B (the marginal rate of substitution of meals for films) shows how the consumer would trade meals for films to maintain a constant level of utility. At point B , the budget line is flatter than the indifference curve. Moves to the left would take the student on to a lower indifference curve because the market trade-off is less than the required utility trade-off.

However, it makes sense to move from B to the right. The market trade-off of affordable meals for films exceeds the utility trade-off required to maintain constant utility. The student reaches a higher indifference curve and increases her utility.

A similar reasoning applies at point E , where the market trade-off of meals for films is less than the utility trade-off needed to hold utility constant. Moves from E to the right reduce utility and take the consumer to a lower indifference curve. Moving from E leftwards along the budget line increases utility and allows the consumer to reach a higher indifference curve.

We can make a general principle out of these examples. Wherever the budget line crosses an indifference curve, a move along the budget line in the smart direction will increase utility. Viewed in these terms, *point C, which maximizes utility, is the point at which the slope of the budget line and the slope of the indifference curve coincide*. Only at point C is there no feasible move along the budget line that increases utility. The student will choose point C since it maximizes utility.

Our analysis implies that a rational consumer chooses a bundle of goods at which the *marginal rate of substitution* (the slope of the indifference curve) equals the *price ratio* (slope of the budget line):⁷

$$MRS = -P_H/P_V$$

To check that our model of consumer choice makes sense, consider what it implies for the observable behaviour of our glutton and film buff whose tastes between meals and films differ. Figure 5.5 represented the indifference curves of the glutton as steep and those of the film buff as flat.

Figure 5.8 assumes these two people have the *same* budget line. They have the same income and face the same prices for food and films. Only their tastes differ. Figure 5.8(a) shows the chosen point C for the glutton, with a lot of meals but few films. The glutton has a strong preference for food (steep indifference curves): her chosen point is far to the right, where the indifference curve flattens out. Figure 5.8(b) confirms that the film buff will choose point C, with many more films but much less food. The film buff has flat indifference curves: her chosen point is far to the left before indifference curves can become flatter than the budget line. Our theory of consumer choice successfully translates differences in taste into observable differences in demand for the two goods.

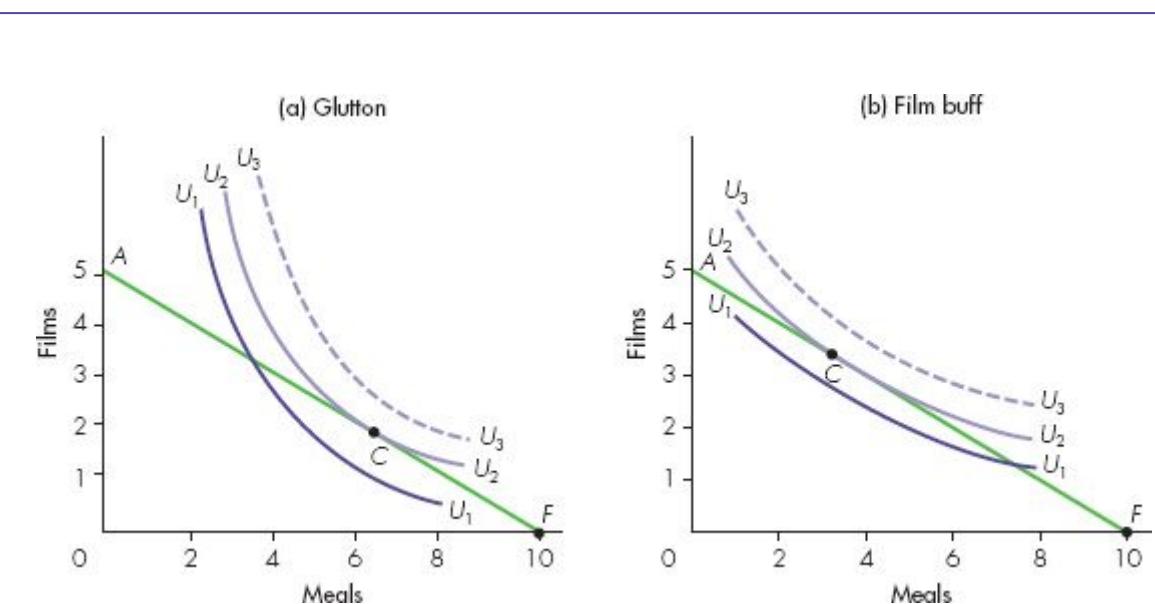


Figure 5.8 The effect of tastes on consumer choice

Both students face the same budget line AF and choose the point C, maximizing utility where the indifference curve is tangent to the budget line. The glutton has steep indifference curves and eats a lot of meals before the diminishing marginal rate of

substitution flattens the indifference curve sufficiently. The film buff has flat indifference curves and the point of tangency is much further to the left. The glutton chooses more meals but fewer films than the film buff.

CONCEPT 5.2

DO CONSUMERS REALLY BEHAVE AS UTILITY-MAXIMIZING AGENTS?

Our model of consumer choice says that consumers, in deciding what to consume, maximize their utility subject to budget constraint. At the chosen consumption bundle, the marginal rate of substitution between the two goods must be equal to their relative price.

This is fine from a theoretical point of view. What about reality? When we go to shops we probably do not write down our indifference curves, calculate our marginal rate of substitution among goods and try to equalize that to relative prices. Probably we do not even know our indifference curves. How can we test our theory, then?

Fortunately, even if we cannot see indifference curves directly, we can indirectly get information about them. Indeed, we can observe consumers' choices. So, can we infer anything about the preferences of a consumer by looking at her consumption choices? The answer is yes. This is the basis of the approach called *revealed preferences*. While a detailed analysis of the revealed preferences approach is beyond the scope of our analysis, we can introduce the basic principle behind it.

The basic idea is to determine consumers' preferences from observing consumers' behaviour. To briefly illustrate this point, suppose that a consumer faces two bundles, X and Y . If she chooses X when Y was also affordable, then we may say that bundle X is revealed as preferred to Y . If our consumer behaves according to our theory, then we should expect her always to choose X instead of Y when both bundles are affordable. If we see our consumer choosing Y instead of X , it should be the case that X has become unaffordable, otherwise our consumer does not behave according to our theory. The important aspect of revealed preferences is that, if

consumer behaviour satisfies some properties (known as the axioms of revealed preferences), then the consumer is indeed a utility-maximizing agent.

The best way to test our theory using revealed preferences is to use experimental data.

We can gather some consumers into a room and ask them to choose among different bundles at given prices. Then we can change the prices and ask them to choose again, and so on. Recent research did just that. One hundred and twenty consumers (randomly selected) from Dijon in France were asked to participate in an experiment in which they had to choose between different bundles in different price/budget configurations. The result of that experiment is that 71 per cent of the consumers indeed behaved as utility-maximizers.

Therefore those consumers, in deciding what bundles to choose, were indeed choosing in such a way that the marginal rate of substitution among goods was equalized to the relative prices, even though they may not have been aware of it.

Source: P. Fevrier and M. Visser, ‘A study of consumer behavior using laboratory data’, *Experimental Economics* 7, no. 1 (2004): 39–114.

5.2 Adjustment to income changes

Chapter 4 introduced the income elasticity of demand to describe, other things equal, the response of quantity demanded to changes in consumer incomes. Now we can use our model of consumer choice to analyse this response in greater detail.

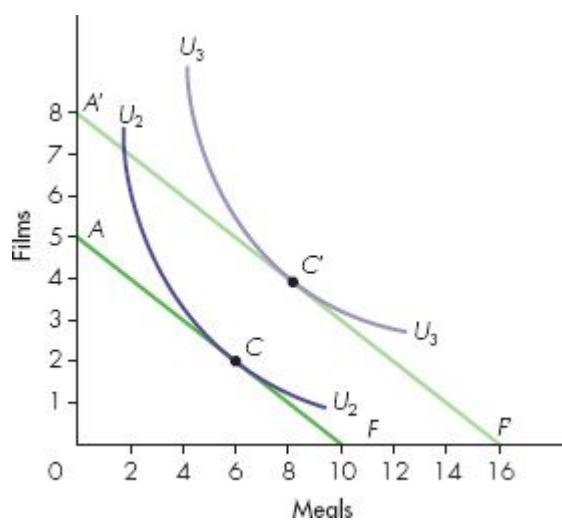
For given tastes and prices, Figure 5.9 shows the effect of a higher income. The student had an income of £50, faced the budget line AF , and chose point C to maximize utility. Suppose her income rises from £50 to £80. Prices of meals and films remain £5 and £10, respectively. With higher income, she can afford to consume more of one or both of the goods. The budget line shifts outwards from AF to $A'F'$.

To find the exact position of this new line, we again calculate the end-points at which all income is spent on a single good. Point A' shows that £80 buys at most 8 films at £10 each. Point F' shows that £80 buys at

most 16 meals at £5 each. Joining these points yields the new budget line $A'F'$. Since the slope of a budget line depends only on the relative price of the two goods, which is unchanged, the new budget line $A'F'$ is parallel to the old budget line AF . Which point on $A'F'$ will the student choose? She chooses C' , at which the new budget line is tangent to the highest attainable indifference curve. However, the position of C' depends on the map of indifference curves that describes her tastes.

For most consumers, food is a normal good but a necessity, whereas films are a luxury. Figure 5.9 shows the case in which the consumer's tastes have these properties. A rise in income from £50 to £80 moves her from C (2 films, 6 meals) to C' (4 films, 8 meals). Thus, a 60 per cent rise in income induces a 100 per cent increase in the quantity of films demanded, confirming that films are a luxury good with income elasticity in excess of unity. Similarly, the 60 per cent rise in income induces a 33 per cent increase in the quantity of meals demanded. The income elasticity of demand for food is $(0.33/0.6) = 0.55$, confirming that food is a normal good (income elasticity greater than zero) but a necessity (income elasticity less than unity).

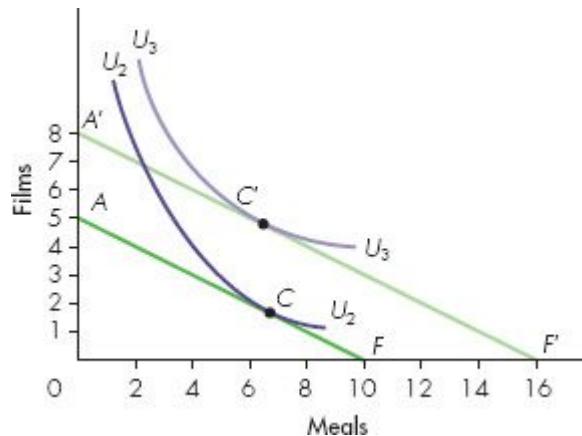
In contrast, in Figure 5.10 her tastes make food an inferior good, for which the quantity demanded declines as income rises. At point C' on the budget line $A'F'$, fewer meals are demanded than at point C on the budget line AF , corresponding to the lower income.



An increase in income from £50 to £80 induces a parallel shift in budget line from AF to $A'F'$. The new end-points A' and F' reflect the increase in purchasing power if only one good is purchased. The slope remains unaltered since prices have not changed. At the higher income the consumer chooses C' . Since both goods are normal, higher income

raises the quantity of each good demanded but the percentage increase in film quantity is larger since its income elasticity is higher.

Figure 5.9 An increase in consumer income



Again, income is increased from £50 to £80 and there is a parallel shift in the budget line from AF to $A'F'$. If meals were an inferior good, the quantity demanded would fall as income rises. The consumer then moves from C to C' when income rises.

Figure 5.10 An increase in income reduces demand for the inferior good

The effects of a fall in income are, of course, exactly the opposite. The budget line shifts inwards but remains parallel to the original budget line. When both goods are normal, lower income reduces the quantity demanded for both goods. If one good is inferior, the quantity demanded will actually rise if income falls. Notice that both goods cannot be inferior: when income falls but prices remain unchanged, it cannot be feasible for the consumer to consume more of both goods.

Income expansion paths

Instead of the response of demand to a particular change in income, we might want to know the response of demand to income in relation to all possible variations in income. To study this, we trace out the **income expansion path** or income-consumption path. Look again at Figure 5.9. The budget lines AF and $A'F'$ correspond to incomes of £50 and £80, respectively. With yet higher incomes we could draw more budget lines, parallel to AF and $A'F'$ but higher up. We could then find the points on

these new budget lines that the consumer would choose at these higher income levels. Joining up the chosen points (C and C' in Figure 5.9) and these new points (say, C'' and C'''), we get the income expansion path. An example of the income expansion path for the case in which both goods are normal is given in Figure 5.11.

The **income expansion path** shows how the chosen bundle of goods varies with consumer income levels, keeping constant everything else.

From the income expansion path we can derive another curve, called the Engel curve.⁸

An Engel curve is just the relationship between the quantity consumed of a good by a consumer and her income, keeping fixed the price of that good. In the case of a normal good, the Engel curve is sloped upward, meaning that as income increases the quantity demanded of that good also increases. In the case of an inferior good, the Engel curve will normally have a backward-bending shape as the demand of an inferior good will decrease as income continues to increase. This is shown in Figure 5.12. In the top graph, we have the income expansion path for two goods, y and x , where good x becomes an inferior good as income increases from M^1 to M^2 . In the bottom graph, we have the Engel curve for good x , which shows that at a low income level the demand for good x increases as income increases, as in the normal good case, but as income continues to increase, good x becomes an inferior good and the consumption of that good starts to decrease.

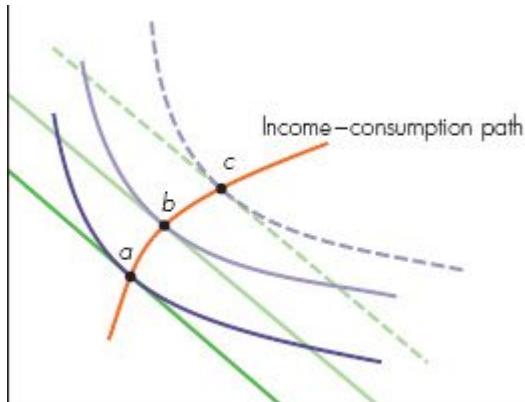


Figure 5.11 Income expansion path for normal goods

5.3

Adjustment to price changes

Having studied changes in tastes and in income, we now isolate the effect of a price change. Chapter 4 argued that a rise in price reduces the quantity demanded, other things equal. The own-price elasticity of demand measures this response, and is larger the easier it is to substitute towards goods whose prices have not risen.

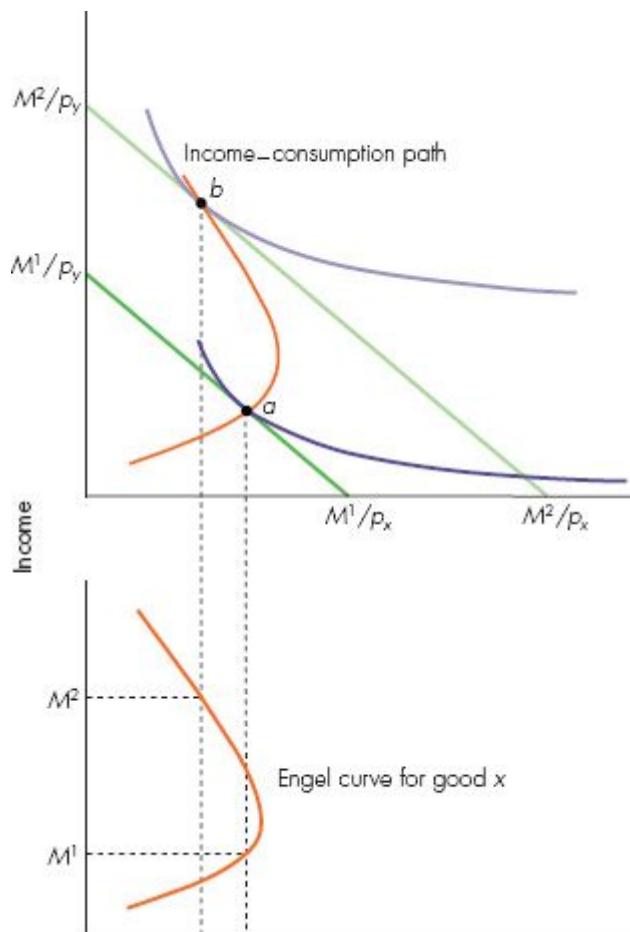


Figure 5.12 Income expansion path and Engel curve for an inferior good

We also introduced the cross-price elasticity of demand to measure the response of the quantity demanded of one good to a change in the price of another good. An increase in the price of good j tends to increase the

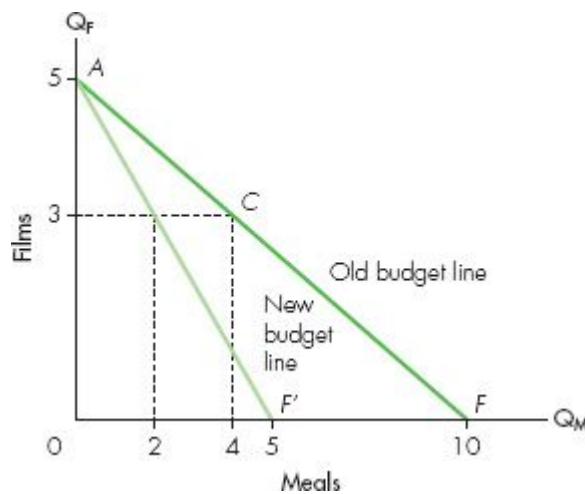
quantity demanded of good i when the two goods are substitutes, but tends to reduce the quantity demanded of good i when the two are complements. The empirical evidence was presented in Tables 4.2 and 4.6.

Are those propositions invariably true, or did the evidence we examined just happen to confirm our commonsense reasoning? We now offer a more formal analysis based on the model of consumer choice developed above.

Price changes and the budget line

Figure 5.13 draws the budget line AF for a consumer with an income of £50 facing prices of £10 and £5 for films and meals, respectively. Suppose meal prices increase to £10. Since the price of films remains unaltered, £50 still buys 10 films when all income is spent on films. Point A must lie on the new budget line as well as the old budget line. But when all income is spent on meals, £50 buys only 5 meals at £10 each, instead of the 10 meals it used to buy at £5 each. Thus the other extreme point on the budget line shifts from F to F' when meal prices double. As usual, we join up these end-points to obtain the new budget line AF' . The effect of a rise in meal prices is to *rotate* the budget line inwards around point A , at which no meals are bought and higher meal prices are irrelevant.

Except at A itself, higher meal prices mean the consumer can now afford fewer meals for any given number of films, or fewer films for any given number of meals. The new budget line AF' lies inside the old budget line AF . The consumption bundles between AF and AF' are no longer affordable at the higher price of meals. In particular, the chosen point on the old budget line is no longer affordable unless it happens to be the end-point A . A price increase makes the consumer worse off by reducing consumption opportunities out of a fixed money income. The consumer's standard of living falls.



The consumer begins at point C on the budget line AF. Doubling meal prices halves the amount that can be spent on meals when no films are bought. The point F shifts to F'. The budget line rotates around the point A at which no meals are bought. Along the new budget line the consumer can no longer afford the original consumption bundle C. Consumption of one or both commodities must be reduced.

Figure 5.13 An increase in meal prices

To check that you understand, try drawing diagrams to illustrate the effect on the budget line of: (1) a reduction in the price of meals (*hint:* Figure 5.13 can be used – how?); (2) an increase in the price of films (*hint:* around which point does the budget line rotate?).

Substitution and income effects

The **substitution effect** of a price change is the adjustment of demand to the relative price change alone.

The **income effect** of a price change is the adjustment of demand to the change in real income alone.

Our model of consumer choice is based on the interaction of affordable opportunities (the budget line) and tastes (indifference curves). To analyse the effect of price changes on the actual quantity of goods demanded, we must study how rotations of the budget line affect the highest indifference curve that the consumer can reach.

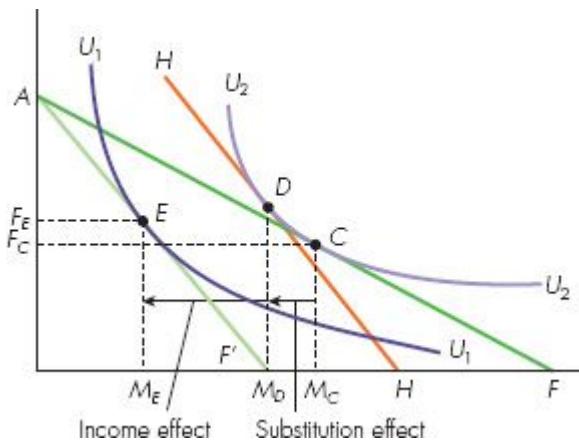
A higher price of meals has two distinct effects on the budget line in Figure 5.13. First, the budget line becomes steeper, reflecting the rise in the relative price of meals. To get an extra meal, more films must now be sacrificed. Second, the budget line AF' lies inside the original budget line AF . The purchasing power of a given money income is reduced by the price increase. If you have £20 and you used to buy chocolate that cost £2 a bar, then your purchasing power is 10 bars of chocolate ($\text{£}20/\text{£}2$). Now suppose that the price of a bar of chocolate rises to £5. Your money now can buy only 4 bars of chocolate.

Economists therefore break up the effect of a price increase into these two distinct effects: the change in the relative price of the two goods and the fall in the purchasing power of the given money income. This is merely a thought experiment but it turns out to be useful to characterize some properties of the goods that consumers choose.

Figure 5.14 shows the response of demand quantities to a higher meal price. At the original prices, the consumer faced the budget line AF and chose C to reach the highest possible indifference curve U_2U_2 . At that point, the consumer demands an amount M_C of meals and an amount F_C of films. If the price of meals increases, the budget line rotates inwards to AF' and the new optimal bundle will be E on the indifference curve U_1U_1 , the highest indifference curve now possible. In this example, higher meal prices reduce the quantity demanded of meals from M_C to M_E , while increasing the quantity demanded of films from F_C to F_E .

The substitution effect

To isolate the effect of relative prices alone, imagine a *hypothetical* budget line HH , parallel to the new budget line (the one after the price of meals has increased, that is, AF') and tangent to the original indifference curve (that is, U_2U_2). Because HH is parallel to the new budget line AF' , its slope reflects the new relative prices of films and meals after the price of meals has risen. Because HH is tangent to the old indifference curve U_2U_2 , it restores the consumer to the original utility and standard of living shown by all points on U_2U_2 . In constructing this new hypothetical budget line, we are doing the following thought experiment: after the price of meals has increased, how much income should we give to our consumer in order for her to have the same level of utility as before?



For a normal good, the income and the substitution effects move in the same direction.

Figure 5.14 Income and substitution effects

This hypothetical income that we can give to our consumer in order to make her as well off as before a price change is what economists call *compensating variation*.⁹

If confronted with the hypothetical budget line HH , the consumer would choose bundle D . Why do we need to do all of this? Because we can now decompose graphically the movement from C to E into two different steps: the movement from C to D and then from D to E .

The movement from C to D depends only on the price change, and we call that the *substitution effect*. The movement from D to E depends only on the fact that the real income has changed, and we call that the *income effect*.

The move from C to D is the pure substitution effect that is the adjustment of demand to relative prices when income is adjusted to maintain the old standard of living in the face of the new higher prices. *The substitution effect of an increase in the price of meals unambiguously reduces the quantity of meals demanded.* This result is perfectly general.¹⁰ As meals become relatively more expensive, the consumer switches towards films, which have become relatively cheaper. Therefore, in moving from C to D , the quantity demanded of meals decreases while the quantity demanded of films increases. In general, the substitution effect is always negative for the good whose price has changed. This means that the consumer will always substitute the good

that is now relatively expensive with the good that is now relatively cheap.

The income effect

To isolate the effect of the reduction in real income, holding relative prices constant, consider now the parallel shift in the budget line from the hypothetical position HH to the actual new position AF' . The consumer moves from D to E . When both goods are normal goods, a reduction in real income will reduce the quantity demanded of both goods. This is the case considered in Figure 5.14, where E lies to the south-west of D . From Figure 5.14 we see that, for meals, the income and substitution effects go in the same direction, meaning they reinforce each other in reducing the quantity consumed of meals. This is a general feature of normal goods. When the price of a normal good changes, the substitution and income effects for that good reinforce each other.

CASE 5.1

INCOME AND SUBSTITUTION EFFECTS IN PRACTICE: THE EFFECTS OF PETROL PRICES ON GROCERY EXPENDITURE

Many consumers use cars to go shopping at grocery shops. Recent research tried to understand how changes in the price of petrol can affect the consumption of grocery products for consumers in California. In order to do that, the researchers used data from the Consumer Expenditure Survey and from detailed scanner data from grocery shops about food products. The research explored the following: suppose that consumers can choose between different bundles containing the following two goods: *food away from home* (like eating out in a restaurant) and *food at home* (like grocery shop food). If the price of petrol increases, eating out and going to the grocery shop using a car become more expensive. How do petrol price increases affect expenditure on those two goods?

The researchers found that, if the price of petrol doubles, the expenditure (and so the consumption) of food away from home decreases by 56 per cent. On the other hand, when the petrol price

doubles, the expenditure on grocery food increases by around 19 per cent. This means that, when the price of petrol increases, food away from home becomes relatively more expensive than food at home, and so the consumers substitute the former with the latter. So the substitution effect works in the same way as we have seen in our analysis.

Another interesting result of this research is that consumers tend to substitute further within their grocery shop purchases when the price of petrol increases. In particular, consumers substitute towards items that are on special offer and away from full-price items when the price of petrol increases substantially.

Source: D. Gicheva et al., ‘Revisiting the income effect: Gasoline price and grocery purchases’, NBER working paper, 2008.

The net effect of a price increase on the quantity demanded

The consumer moves directly from the original point C to the new point E . We can interpret this as a pure substitution effect from C to the hypothetical point D , plus a pure income effect from D to E . If the good whose price has risen is a normal good, demand curves slope downwards, as asserted in Chapter 4.

The substitution effect from C to D must reduce the quantity of meals demanded. When the price of meals rises, the budget line becomes steeper and we must move along U_2U_2 to the left to find the point at which it is tangent to HH . Similarly, the income effect must further reduce the quantity of meals demanded if meals are a normal good. E must lie to the left of D .

The individual demand curve

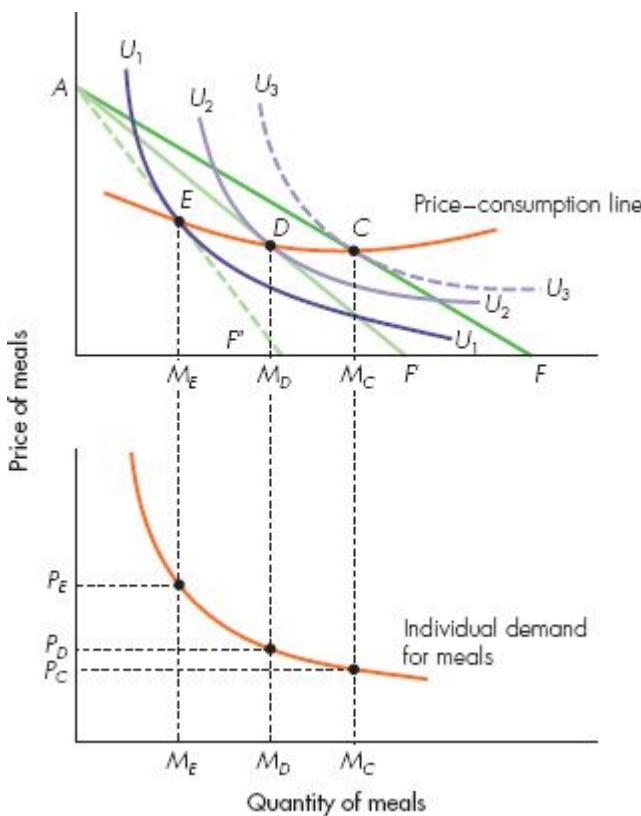
Now we have all the tools we need to derive demand by a single consumer for a given good. In Figure 5.15 we show how to derive graphically the individual demand for a normal good. The top part shows how the optimal choice of our consumer changes as the price of meals increases, everything else constant. Suppose we start at point C , where the price of meals is P_C . At that price, the consumer demands an amount of meals given by M_C . Now suppose that the price increases to P_D . The

budget constraint rotates inwards to AF' ; at the new price of meals, the optimal choice of the consumer is now bundle D . So at price P_D our consumer demands an amount M_D of meals.

Then suppose that the price of meals increases even further, say to P_E . At this new price the optimal choice is point E and our consumer demands an amount M_E of meals. The line joining all the optimal bundles is called the *price-consumption curve*.

The bottom part of Figure 5.15 shows a graph with the price of meals on the vertical axis and the quantity consumed (and so demanded) of meals on the horizontal axis. Using the information in the top graph we can derive a possible negative relationship between the price of meals and the quantity demanded of meals. This is the individual demand for meals of our consumer. In the case where the good is normal, the demand curve implies a negative relationship between the quantity demanded of that good and its market price.

Our analysis also explains what determines a consumer's willingness to pay for a given good. It is determined by her preferences and her income (as described by the budget line). Two consumers with the same preferences for a good may exhibit different willingness to pay for it because they have different incomes.



The individual demand for meals can be derived from the solution of the consumer problem once we allow the price of meals to vary keeping constant everything else.

Figure 5.15 Income and substitution effects

Inferior goods

Although the substitution effect must reduce the quantity of meals demanded when the price of meals increases, the income effect goes in the opposite direction if the good is inferior: reductions in real income increase the quantity demanded. We can even imagine a perverse case in which this effect is so strong that price rises actually increase the quantity of that good demanded. Demand curves then slope *upwards!* We can use our analysis to explain how such a paradoxical case can arise.

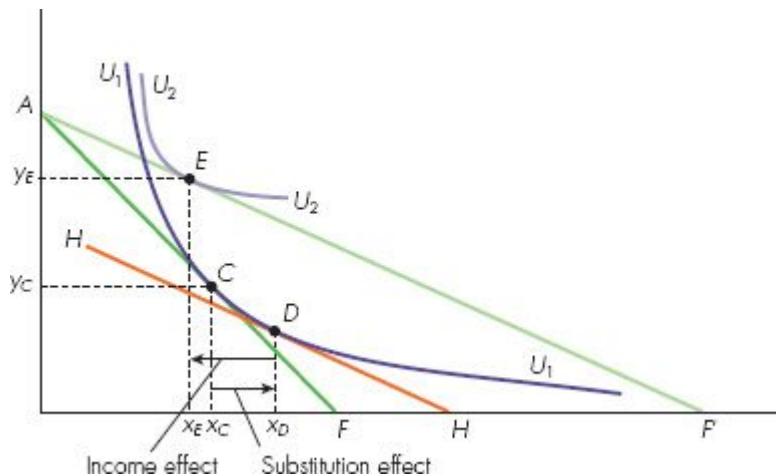
Consider the case of two different goods, X and Y . This is represented in Figure 5.16. Suppose that the price of good X decreases. As usual after a price change, we can decompose the effect of this price change into substitution and income effects. Suppose that, before the decrease in the price of X , the optimal choice of the consumer was bundle C on the

indifference curve U_1U_1 . After the price of X decreases, the budget constraint rotates outwards from AF to AF' . The new choice of the consumer is bundle E on the indifference curve U_2U_2 . To show the income and substitution effects, we draw a new budget line HH parallel to AF' and tangent to the original indifference curve (U_1U_1). By doing this, we identify bundle D . The movement from C to D is due to the substitution effect. Since the price of X decreased, good X is now relatively cheaper than good Y and so the consumer substitutes Y with X . The movement from D to E is due to the income effect. As income increases, the quantity demanded of good X decreases (from x_D to x_E). In this case, X is an inferior good. For an inferior good, it is always true that income and substitution effects go in opposite directions. In the particular case considered in Figure 5.16, the income effect is larger than the substitution effect and the final result is that, after the price of X has decreased, the quantity demanded of X has decreased (from x_C to x_E). In this case, the demand curve of good X is upward sloping.

A good that has such a property (the income effect is bigger than the substitution effect) is called a Giffen good, after a nineteenth-century economist who examined whether higher potato prices raised the quantity of potatoes demanded by the poor.

Notice in Figure 5.16 that, while good X is a Giffen good, Y is instead a normal good. As income increases, the quantity demanded of Y increases.

For a Giffen good, the demand curve is upward sloping. However an inferior good need not be a Giffen good. It requires a very strong income effect – here, an increase in demand in response to real income reductions – to offset the substitution effect that is always negative. When goods are inferior but not Giffen, their demand curve is downward sloping, as shown in Figure 5.15. After decades of empirical research, economists are convinced that Giffen goods are rare. In practice, goods are rarely so inferior that the income effect can reverse the substitution effect. This means that, apart from the rare cases given by Giffen goods, for almost all goods we should have a negative demand curve.



A Giffen good is an inferior good for which the income effect is positive and larger than the substitution effect (always negative).

Figure 5.16 A Giffen good

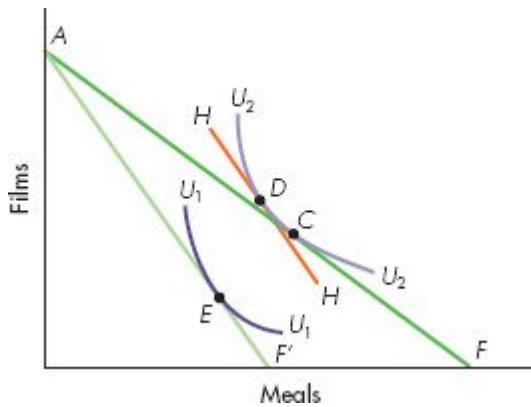
Cross-price elasticities of demand

How does a rise in the price of one good affect demand for other goods? Chapter 4 showed that cross-price elasticities may be negative or positive. We now illustrate these possibilities, highlighting the roles played by substitution and income effects.

Figure 5.17 shows a negative cross-price elasticity. A higher price of meals reduces the quantity of films demanded. Figure 5.17 has three properties. First, the two goods are poor substitutes. Indifference curves are very curved. Moving away from balanced combinations of the two goods requires large extra quantities of one good to compensate for small losses of the other good if a constant level of utility is to be preserved. When the price of meals is increased, the substitution effect towards films is small. Moving leftwards along U_2U_2' , we quickly attain the slope required to match the new relative prices of the two goods. The substitution effect from C to D adds little to the quantity of films demanded.

Second, films have a high income elasticity of demand. They are a luxury good. Hence the income effect, the move from D to E in response to the parallel downward shift in the budget line from HH to AF' , leads to a lot fewer films demanded.

Finally, point C is well to the right on the original budget line AF . Meal expenditure is a large part of consumer budgets. Hence changes in meal prices lead to big changes in the purchasing power of consumer income. Not only is the number of films demanded very responsive to given changes in consumer real income, but also a given rise in meal prices has a large effect on consumer real income because meals are a large part of consumer budgets.



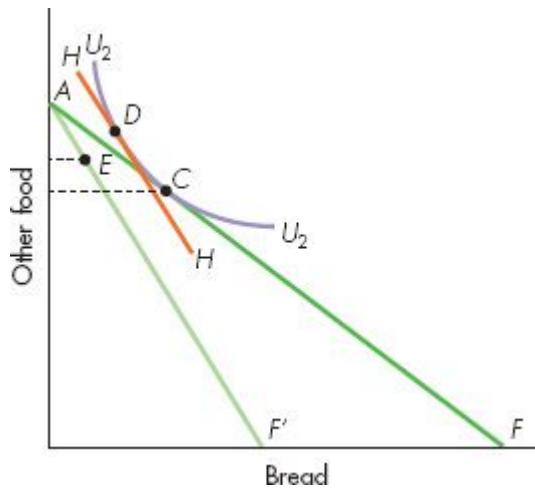
An increase in the price of meals rotates the budget line from AF to AF' . The substitution effect from C to D is small. Indifference curves have large curvature since the two goods are poor substitutes in utility terms. The income effect from D to E implies a large reduction in films for two reasons. First, the reduction in real income is larger the further to the right the initial point C . Second, films are a luxury good whose quantity demanded is sensitive to changes in real income. Thus the income effect outweighs the substitution effect. E lies below C .

Figure 5.17 A negative cross-price elasticity

These last two effects lead to a large income effect, which reduces the quantity of films demanded. Because the substitution effect in favour of films is small, the net effect is a reduction in the quantity of films demanded. An increase in meal prices reduces the quantity of films demanded. The cross-price elasticity of demand is negative.

Figure 5.18 shows the opposite case, a positive cross-price elasticity of demand. Suppose the consumer is choosing between bread and other food. If the price of bread rises, potatoes are a good substitute for bread. To maintain a given utility, consumers can substitute lots of cheap potatoes for expensive bread. Indifference curves are less curved than in Figure 5.17.

Suppose also that other food has a small income elasticity of demand. Although higher bread prices reduce real consumer income, this has a small income effect that reduces the quantity of other food demanded. Finally, if bread is a relatively small share in consumer budgets, higher bread prices have a small effect in reducing consumer purchasing power. Comparing Figures 5.17 and 5.18, the parallel shift from HH to AF' is smaller in the latter.



An increase in the price of bread rotates the budget line from AF to AF' . The substitution effect from C to D is large. Indifference curves have little curvature since the two goods are good substitutes in utility terms. The income effect from D to E is relatively small because the income elasticity of demand for other food is low and because the reduction in real income is small since bread forms a small share of the consumer budget. The substitution effect outweighs the income effect. E lies above C .

Figure 5.18 A positive cross-price elasticity

These last two effects imply that there is only a small income effect reducing the quantity of other food demanded. In contrast, the substitution effect towards other food is big. Hence higher bread prices raise the quantity of other food demanded. The cross-price elasticity is positive. This positive effect is even stronger if ‘other food’ is an inferior good. The income effect then raises the quantity of other food demanded, reinforcing the substitution effect. Table 5.2 summarizes the implications of our model of consumer choice for the demand response to a price change.

Table 5.2 The effect of an increase in the price of good I on the quantity demanded of goods I and J

Good	Type	Substitution effect	Income effect	Total effect
I	Normal	Negative	Negative	Negative
	Inferior	Negative	Positive	Ambiguous
J	Normal	Positive	Negative	Ambiguous
	Inferior	Positive	Positive	Positive

For example, suppose good *I* and good *J* are both normal. If the price of good *I* increases, we know that the consumer will substitute good *I* with good *J*, which is now relatively cheaper. So the substitution effect has a negative impact on the demand for good *I*, while it has a positive impact on the quantity demanded of good *J*. The increase in the price of good *I* decreases the real income of the consumer and, since goods are normal, the income effect has a negative impact on the quantity demanded of both goods.

CASE 5.2

ARE CIGARETTES INFERIOR GOODS?

A recent phone survey about smoking behaviour in the US provided the results displayed in the following table.

Do you smoke? Percentage 'yes' among Americans aged 30 to 64					
		Lower household income		Higher household income	
		Less than \$24 000 (%)	\$24 000 to less than \$36 000 (%)	\$36 000 to less than \$90 000 (%)	\$90 000 and more (%)
Lower education	Less than high school	42	36	40	40
	High school graduate	39	32	26	23
	Some university/Vocational school	38	29	21	18
	University graduate/Postgrad	22	15	10	7
Higher education					

Americans aged from 30 to 64 were asked whether or not they smoke, their income and their level of education. If we look at the results of the survey, we see that people with lower incomes (less

than \$24 000) make up the majority (42 per cent) of smokers. Moreover, as income increases, fewer and fewer people tend to smoke. This is true for any level of education. The relationship between number of smokers and income level seems to suggest that cigarettes are inferior goods. We know that the demand for cigarettes tends to be negatively sloped, so cigarettes are inferior but not a Giffen good. Why is that?

One possible reason is related to level of education. Better-educated people have greater access to information on the severe health problems resulting from smoking. They are also more likely to accept the truth of such information. Moreover, level of education is normally positively related to income. Those two facts together can explain why cigarettes may be an inferior good.



©Konstantin Tavrov | Dreamstime.com

Source: ‘Smoking is an inferior good (sometimes)’, *The Economist*, 29 April 2010. © The Economist Newspaper Limited, London 2010.

5.4 The market demand curve

We have now established that individual demand curves (almost always) slope downwards. For the rest of this book we assume that this is the case. Once we have the individual demand for each consumer demanding a given good, we can find the market demand for that good. We get the market demand by aggregating the demand curves of every individual consumer buying a particular good.

Consider the simplest case where there are only two consumers buying a given good. At each price, we find out how much each consumer

demands. Adding the quantities demanded by all consumers at that price, we get the total quantity demanded at each price – the market demand curve. Since, as price is reduced, each person increases the quantity demanded, the total quantity demanded must also increase as price falls. The market demand curve also slopes downwards.

The **market demand curve** is the sum of the demand curves of all individuals in a particular market.

The **market demand curve** is the *horizontal addition of individual demand curves*. With prices on the vertical axis and quantities on the horizontal axis, we must add together individual quantities demanded at the same price. Figure 5.19 illustrates this idea.

Suppose that, when the price is £3, consumer 1 demands 4, while consumer 2 demands 12. Then, at a price of £3, the market quantity demanded is $4 + 12 = 16$. This is one point of the market demand. Suppose that, when the price is £5, consumer 1 demands 0, while consumer 2 demands 10, then at $p = £5$, the market quantity is 10. This is a second point of the market demand. Notice that, in the particular example in Figure 5.19, when the price is above £5, consumer 1 always demands 0. Therefore the market demand coincides with the demand of consumer 2 at prices above £5. At prices below £5, the market demand is the horizontal sum of the quantities demanded by each consumer at the same price.

5.5 Complements and substitutes

Income and substitution effects are used to understand the effects of a price change. Whatever the direction of the income effect, with only two goods the substitution effect is always negative. The pure relative price effect leads the consumer to substitute away from the good whose relative price has risen towards the good whose relative price has fallen. Abstracting from income effects, goods are necessarily substitutes for one another in a two-good world.

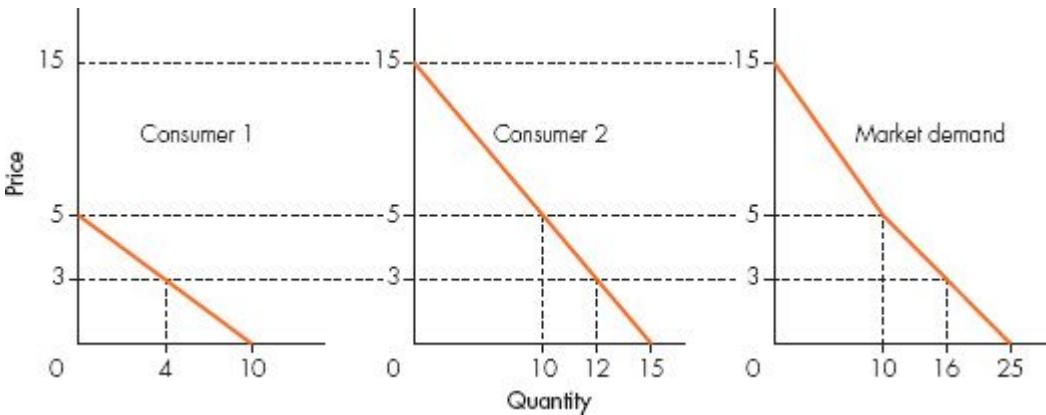


Figure 5.19 Individual demand curves and the market demand curve

With more than two goods, some goods may be consumed jointly – pipes and pipe tobacco, bread and cheese, electric cookers and electricity. These goods are *complements*.

Even with many goods, there is always a substitution effect *away* from goods whose relative price has risen. However, substitution may not be *towards* all other goods. Consumers substitute *away* from goods consumed jointly with the good whose price has risen.

Suppose the price of pipes rises. What will happen to the demand for pipe tobacco? (Ignore the income effect, since expenditure on pipes is a tiny fraction of household budgets, so real incomes are only slightly reduced.) Since pipes and pipe tobacco are used jointly, we expect the demand for pipe tobacco to fall along with the number of pipes demanded. The demand curve for pipe tobacco shifts to the left in response to the increase in pipe prices. Notice that this implies that the cross-price elasticity between those two goods should be negative.

When goods are complements, a rise in the price of one good will reduce the demand for the complement, both through the substitution effect (substituting away from the higher-priced activity) and, of course, through the income effect (provided goods are normal).

Special cases are goods that are *perfect substitutes* or *perfect complements*. Perfect substitutes are goods that are viewed as equal by consumers. In that case, the consumer always consumes the cheap one. Suppose that, for you, Coca-Cola and Pepsi are exactly the same thing. Then you will drink the one that costs less. On the other hand, perfect

complements are goods that are always consumed together in fixed proportion, for example one right shoe and one left shoe.

The indifference curves in these two particular cases are displayed in Figure 5.20.

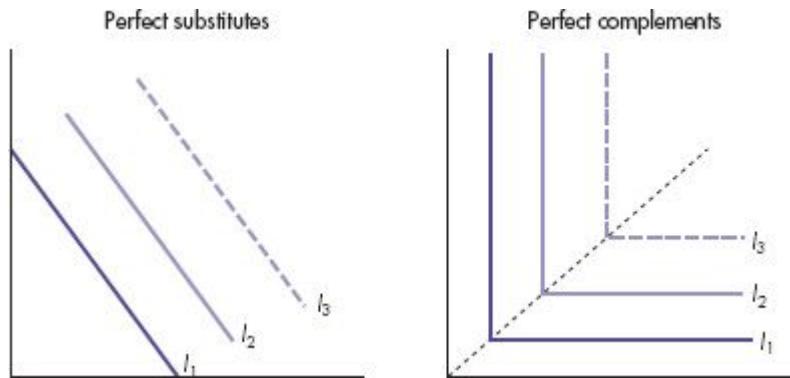


Figure 5.20 Indifference curves for perfect substitutes and complements

Higher indifference curves imply higher utility. In the case of perfect substitutes, the indifference curves are downward-sloping straight lines. Therefore, for perfect substitutes the marginal rate of substitution is constant. In this particular case, at the optimal choice it is not normally true that the marginal rate of substitution is equal to the slope of the indifference curve. We have a corner solution. Our consumer consumes only the cheapest good.

The case of perfect complements is one where the property that consumers prefer ‘more to less’ does not hold. Suppose you really like to eat your chocolate cake with a scoop of vanilla ice cream on top. You prefer that to the same cake with two scoops of ice cream. Having more of one of the two goods does not make you happier, since you always prefer to consume the two goods in fixed proportion. This explains the L-shaped indifference curves. The slope of the dashed straight line starting from the origin measures the fixed proportion in which the two goods are consumed.

5.6 Transfers in kind

Social security payments are a monetary **transfer**. Wages are not: the recipient provides labour services in exchange for wages. An example of a **transfer in kind** is food stamps, given to the poor to buy food. The stamps must be spent on food, not beer, films or petrol. We now use our model of consumer choice to ask whether an in-kind transfer payment is preferred by the consumer to a cash transfer payment of the same monetary value.

A **transfer payment** is a payment, usually made by the government, for which no corresponding service is provided by the recipient.

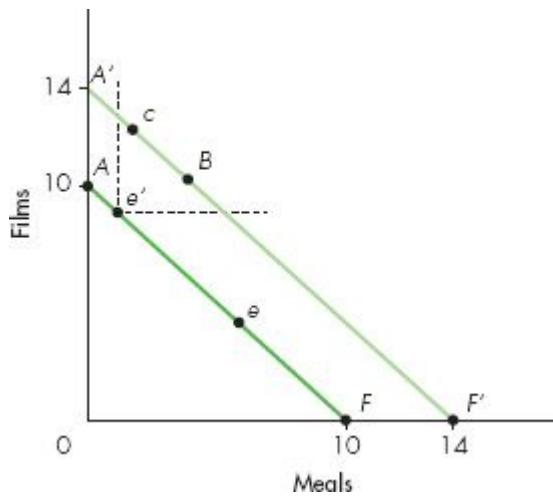
A **transfer in kind** is the gift of a good or service.

The consumer has £100 to spend on food or films, each costing £10 per unit. Figure 5.21 shows the budget line AF . Suppose the government issues the consumer with stamps worth 4 food units. For any point on the old budget line AF , the consumer can have 4 more units of food from the food stamps. Moving horizontally to the right by 4 food units and given that food stamps cannot buy films, the new budget line is ABF' . The consumer can still get at most 10 films.

Suppose the consumer originally chose e on the budget line AF . Since both goods are normal, the shift in the budget line to ABF' – effectively a rise in income – makes the consumer choose a point to the north-east of e , as she would have done had the transfer been in cash.

When food costs £10 per unit, the cash equivalent of 4 food units is £40, shifting the budget line to $A'F'$. Thus, if the consumer begins at e , it makes no difference if the transfer is in cash or in kind.

Suppose, however, that the consumer begins at e' . With a cash payment, the consumer might move to point c on the budget line $A'F'$. The transfer in kind, by restricting the consumer to the budget line ABF' , prevents her reaching the preferred point c . Instead she moves, say, to the feasible point B . B must yield the consumer less utility than c : when she got a cash payment and could choose either point, c was preferred to B .



A food transfer in kind may leave consumers less satisfied than a cash transfer of the same value. A consumer at e' might wish to spend less than the full allowance on food, moving to c . The budget line is $A'B'F'$ under a cash transfer. The in-kind transfer restricts the budget line to ABF' .

Figure 5.21 Transfers in cash and in kind

Cash transfers let consumers spend the extra income in any way they wish. Transfers in kind may limit a consumer's options. Where they do, the increase in consumer utility is less than under a cash transfer of the same monetary value.

Yet transfers in kind are politically popular. The electorate wants to know that taxes are being wisely spent. Some people argue that the low-income people may not know how to spend their money wisely and may spend cash transfers on 'undesirable' goods such as alcohol or gambling rather than on 'desirable' goods such as food or housing.

Do people act in their own best interests? This issue is not merely one of economics but also of philosophy, involving wider questions such as liberty and paternalism. In so far as people can judge their own self-interest, economic analysis is clear: people are better off, or at least no worse off, if they get transfers in cash rather than in kind.

Summary

- Given the **budget constraint**, the theory of demand assumes a consumer seeks to reach the **maximum possible level of utility**.
- The **budget line** shows the maximum affordable quantity of one good for each given quantity of the other good. The position of the budget line is determined by income and prices alone. Its slope reflects only relative prices.
- Because the **consumer prefers more to less**, she will always select a point on the budget line. The consumer has a problem of choice. Along the budget line, more of one good can be obtained only by sacrificing some of the other good.
- **Consumer tastes** can be represented by a map of non-intersecting **indifference curves**. Along each indifference curve, utility is constant. Higher indifference curves are preferred to lower indifference curves. Since the consumer prefers more to less, indifference curves must slope downwards. To preserve a given level of utility, increases in the quantity of one good must be offset by reductions in the quantity of the other good.
- Indifference curves exhibit a **diminishing marginal rate of substitution**. Their slope is flatter as we move along them to the right. To maintain given utility, consumers sacrifice ever-smaller amounts of one good to get successive unit increases in the amount of the other good.
- **Utility-maximizing** consumers choose the consumption bundle at which the highest reachable **indifference curve is tangent to the budget line**. At this point, the market trade-off between goods, the slope of the budget line, just matches the utility trade-off between goods, the slope of the indifference curve.
- At constant prices, an increase in income leads to a parallel outward shift in the budget line. If goods are normal, the quantity demanded will increase.

- A change in the price of one good rotates the budget line around the point at which none of that good is purchased. Such a price change has an income effect and a substitution effect. The **income effect of a price increase** is to reduce the quantity demanded for all normal goods. The **substitution effect**, induced by relative price movements alone, leads consumers to substitute away from the good whose relative price has increased.
- In a two-good world, goods must be substitutes. The substitution effect is unambiguous. With many goods, the pure substitution effect of a price increase also reduces demand for goods that are complementary to the good whose price has risen.
- A rise in the price of a normal good must lower its quantity demanded. For inferior goods, the income effect operates in the opposite direction but rarely seems to dominate the substitution effect. Demand curves slope downwards.
- The **market demand curve** is the horizontal sum of individual demand curves, at each price adding together the individual quantities demanded.
- **Consumers prefer to receive transfers in cash rather than in kind**, if the two transfers have the same monetary value. A transfer in kind may restrict the choices a consumer can make.

Review questions



EASY

- 1 **True or False** On a given indifference curve, the marginal rate of substitution is *always* decreasing.
- 2 **Common fallacies** Why are these statements wrong? (a) Since consumers do not know about indifference curves or budget lines, they cannot choose the point on the budget line tangent to the highest possible indifference curve. (b) Inflation must reduce demand since prices are higher and goods are more expensive.

- 3 Suppose films are normal goods but transport is an inferior good. How do the quantities demanded for the two goods change when income increases?
- 4 The own-price elasticity of demand for food is negative. The demand for food is inelastic. A higher food price increases spending on food. Higher food prices imply less is spent on all other goods. The quantity demanded of each of these other goods falls. Discuss each statement. Are they all correct?
- 5 Suppose Glaswegians have a given income and like weekend trips to the Highlands, which are a three-hour drive away. (a) If the price of petrol doubles, what is the effect on the demand for trips to the Highlands? Discuss both income and substitution effects. (b) What happens to the demand for Highland hotel rooms?

MEDIUM

- 6 Frank's utility function for two goods, X and Y , is given by $U = XY$. Find Frank's indifference curves, when utility is 10, 20 and 30. Plot these indifference curves. How should Frank compare the following two bundles: $(X = 1, Y = 10)$ and $(X = 5, Y = 2)$?
- 7 Suppose Frank has an income of £50, the unit price of X is $P_X = \text{£}2$ and the unit price of Y is $P_Y = \text{£}1$. Write down the budget constraint for Frank. Knowing that the marginal rate of substitution (in absolute value) between X and Y is $MRS = X/Y$, find the optimal bundle that Frank should consume. (*Hint:* at the optimal bundle, the absolute value of the MRS must be equal to the absolute value of the slope of the budget constraint. Moreover, the budget constraint must be satisfied. You need to solve a system of two equations in two variables, X and Y .)
- 8 Suppose you have 5 coconuts and 5 fish. You can get extra fish by sacrificing 2 coconuts for each extra fish, or get extra coconuts by sacrificing 1 extra fish for each extra coconut. (a) Draw your budget line. (b) Draw an indifference map. (c) Where is it likely that you will choose to be? (d) Suppose there is a small change in the number of fish you can swap for an extra coconut – is your behaviour likely to change?
- 9 Refer to the graph in Figure 5.12 (page 97) to answer the question. The graph shows the income consumption path for goods X and Y , and the Engel curve for good X .
- (a) What does the backward-bending income consumption path imply about good Y ?

- (b) Given the Engel curve in the graph, identify the nature of good X .
- |0 You can invest in a safe asset, in a risky asset, or in both. The safe asset has a guaranteed return of 3 per cent a year. The risky asset has an expected return of 4 per cent but it could be as much as 8 per cent or as little as 0 per cent. You decide to have some of your wealth in each asset. Now the expected return on the risky asset rises to 5 per cent; it could be as high as 9 per cent or as low as 1 per cent. Given the increase in the expected return on the risky asset, do you invest more of your wealth in the risky asset?
- |1 Suppose that Carl cannot tell the differences between a pack of British and a pack of Danish bacon. In a graph with British bacon on the vertical axis, plot some of Carl's indifference curves for British and Danish bacon. Suppose that Carl has an income of £20. The price of Danish bacon is £2 per pack, while the price of British bacon is £4 per pack. Using the same graph, draw Carl's budget constraint and show his optimal bundle choice.

HARD

- |2 **Essay question** Consumer choice theory assumes that consumers are rational but we observe a person behaving differently in apparently similar situations. Is it realistic to think that we account for rational behaviour in every situation?
- |3 A consumer's income is £50. Food costs £5 per unit and films cost £2 per unit. (a) Draw the budget line. Pick a point e as the chosen initial consumption bundle. (b) The price of food falls to £2.50. Draw the new budget line. If both the goods are normal, what happens to consumption? (c) The price of films also falls to £1. Draw the new budget line and show the chosen point e'' . (d) How does e'' differ from e ? Why?
- |4 Consider a consumer who consumes only two goods, peas and beans. He has an income of £10, the price of beans is 20p per kg (= £0.2) and the price of peas is 40p per kg (= £0.4).
(a) Suppose that the consumer consumes 30 kg of beans. Assuming that the consumer wants to spend all his income, how many kg of peas is he going to consume?
(b) Assume that the price of peas falls from 40p to 20p. Assuming that the consumer still consumes 30 kg of beans, find the new quantity of peas.

- (c) After the decrease in the price of peas to 20p, assume that the consumer is just as well off as he was in (a) if he has an income of £7.60. However, with that income and the new price of peas he would have consumed 20 kg of beans. Find the quantity of peas he would have consumed in this case.
- (d) Find the substitution effect on consumption of peas due to the decrease in the price of peas in (c).
- (e) Find the income effect on consumption of peas due to the decrease in income in (c).

-
- 1 We say that utility is an *ordinal* measure. The case of measurable utility, or *cardinal* utility, is considered in the Appendix to this chapter. This is a special case, but easier to master. Also in the Appendix we analyse more formally the relationship between ordinal utility and preferences.
 - 2 The main results we get from our analysis will not change if we allow consumption bundles to contain more than two goods. However, in that case, the analysis will be slightly more complicated.
 - 3 In the Appendix of this chapter we will show how the marginal rate of substitution is related to the concept of marginal utility.
 - 4 Here, we assume that the consumer has a given income to spend. We do not investigate where this income comes from, whether from working or other sources.
 - 5 In most markets consumers accept the prices offered in stores and their purchases do not affect those prices. In Chapter 10 we will consider the case where a buyer is not a price-taker, a case called monopsony.
 - 6 We assume that all income is spent. There is no saving. In Chapter 11 we discuss the choice between spending and saving.
 - 7 Notice that the *MRS* is negative since an indifference curve is downward sloping.
 - 8 Ernst Engel was a nineteenth-century German statistician who showed how household expenditure on a particular good changes as the household's income changes. In particular, he showed that poorer households tend to spend a larger share of their income in food consumption, something that is now known as Engel's Law.
 - 9 In the case of a decrease in the price of meals, the compensating variation would be the amount of income that we need to take away from our consumer in order for her to be as well off as before.
 - 10 With only two goods, substitution away from meals must imply substitution towards films. However, when there are more than two goods, we cannot be sure that the substitution effects will tend to increase the quantity demanded for all other goods. We discuss this shortly in Section 5.5.

Appendix

Consumer choice with measurable utility

Our theory of consumer choice assumed that consumers can rank different bundles according to the utility or satisfaction they give. Saying bundle A gives more utility than bundle B just means the consumer prefers A to B . We do not need to know *by how much* A is preferred to B . Higher indifference curves are better. We do not need to know how much better.

Nineteenth-century economists believed utility levels could actually be measured, as if each consumer had a *utility meter* measuring his happiness. The further to the right the needle on his utility meter, the happier he was. The units on this meter were traditionally marked off in *utils*. Nowadays, this seems a bit strange: are you 2.9 times as happy if you get an extra week's holiday?

Even so, analysis of consumer choice when utility *is* measurable is quite interesting, even though we derived all the main propositions in the text without this extra assumption. The (robot-like) individual whose utility is exactly calibrated in utils we shall call Fred.

Fred goes to rock concerts and eats hamburgers. For a given consumption of one of these goods, he prefers more of the other to less. His utility goes up. If Fred gets 67 utils of utility from consuming 10 hamburgers and 1 rock concert, and 70 utils from 11 hamburgers but still 1 rock concert, his **marginal utility** from the eleventh hamburger is $(70 - 67) = 3$ utils.

Fred was not very hungry. He had 10 hamburgers at his only concert. He did not get much from an eleventh hamburger, only an extra 3 utils. In contrast, if Fred had only 2 hamburgers at 1 concert (giving him, say, 20 utils), he might rather have enjoyed a third hamburger (taking his utils to, say, 27). The marginal utility of that extra hamburger is $(27 - 20) = 7$ utils. Fred's tastes obey the law of **diminishing marginal utility**.

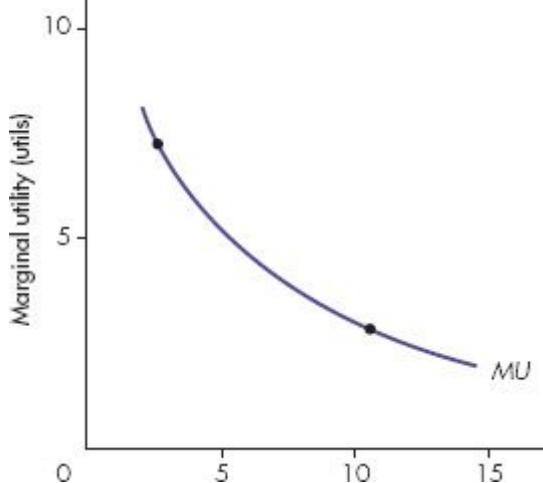
The **marginal utility** of a good is the increase in total utility obtained by consuming one more unit of that good, for given consumption of other goods.

A consumer has **diminishing marginal utility** from a good if each extra unit consumed, holding constant consumption of other goods, adds successively less to total utility.

Figure 5.A1 plots Fred's marginal utility of hamburgers. He gets fewer *extra* utils from extra consumption of hamburgers, the more he is already consuming: his marginal utility schedule MU slopes down.

Fred has a given income to spend. Once we know the prices of rock concerts and hamburgers, we can work out his budget line. How does Fred choose the affordable point on this line at which to consume? He maximizes his utility.

The price of hamburgers in pounds is P_H and the price of concerts is P_C . If MU_H is Fred's marginal utility from another hamburger, he gets an extra MU_H/P_H utils for each extra pound spent on hamburgers and an extra MU_C/P_C utils for each extra pound spent on concerts.



MU shows marginal utility, the amount by which total utils increase when consumption increases one unit. Diminishing marginal utility means that MU falls as quantity rises.

Figure 5.A1 Marginal utility

Suppose MU_H/P_H exceeds MU_C/P_C . An extra pound spent on hamburgers raises Fred's utility more than does an extra pound spent on concerts. If Fred spends £1 more on hamburgers but £1 less on concerts, his total utils rise: he gains more from hamburgers than he loses from

concerts. He can increase utility *without spending more*. He will always want to transfer spending towards the good that yields more marginal utility per pound spent. To maximize utility, Fred spends all his income (he is on, not inside, his budget line) and adjusts his spending between hamburgers and concerts until

$$MU_H/P_H = MU_C/P_C \quad (A1)$$

When this holds, Fred cannot rearrange the division of his total spending to increase his utility. Fred maximizes utility by choosing the consumption bundle, on the budget line, at which the ratio of marginal utility to price is the same for every good.

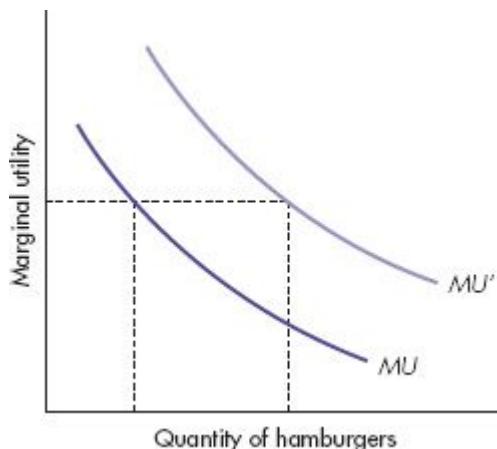
Equation (A1) implies $MU_H/MU_C = P_H/P_C$. Multiplying both sides by -1 , the right-hand side is the slope of the budget line, which depends only on relative prices. The left-hand side is the marginal rate of substitution: if the marginal utility of one hamburger is 2 and of one concert is 4, then $-MU_H/MU_C = -1/2$. One hamburger can be swapped for $1/2$ a concert without altering total utility, precisely what the marginal rate of substitution measures.¹¹ Equation (A1) implies that the slope of the indifference curve, the marginal rate of substitution, equals the slope of the budget line. This is the tangency condition described in the text, derived without using measurable utility.

Deriving demand curves

Suppose the price of hamburgers P_H falls. For given hamburger consumption, MU_H/P_H rises, and now exceeds MU_C/P_C for concerts, violating equation (A1). To maximize utility, Fred changes the quantities he demands.

If Fred buys *more* hamburgers when the price *falls*, the law of diminishing marginal utility means that MU_H falls as Fred buys more hamburgers. MU_H/P_H moves towards MU_C/P_C as required by equation (A1). This is the *substitution effect* of the relative change in the price of hamburgers and concerts. On its own, the substitution effect suggests that *demand curves slope down*: when the price of hamburgers falls, the quantity demanded increases.

However, cheaper hamburger prices also raise the purchasing power of Fred's money income. This affects Fred's marginal utility. If hamburgers are a normal good, Fred buys more when the purchasing power of his income rises. Higher income shifts Fred's marginal utility schedule up in Figure 5.A2.



If hamburgers are a normal good, higher income shifts the marginal utility schedule upwards. The quantity demanded must increase if marginal utility is to remain unaltered.

Figure 5.A2 An increase in the purchasing power of income

CONCEPT A5.1

MARGINAL UTILITY AND THE WATER–DIAMOND PARADOX

Nineteenth-century economists wondered why the price of water, essential for survival, was so much lower than that for decorative diamonds. One answer is that diamonds are scarcer than water. Yet consumers clearly get more total utility from water (without it, they die) than from diamonds. The concept of marginal utility solves the problem.

Equation (A1) tells us that consumers keep buying a good until the ratio of its *marginal* utility to price equals that for other goods. *At the margin*, the last litre of water we drink or use in the shower gives very little extra utility. At the margin, the last diamond still makes a big difference. People are willing to pay more for extra diamonds than for extra water.

In terms of a figure like Figure 5.A1, the marginal utility schedule MU is *very* high for the first few drops of water. Not dying is worth lots of utils. But most of us are a long way down this schedule, using lots of water to the point where its marginal value to us is low.

This *income effect* means that Fred finds that MU_H / P_H rises not only because P_H falls but also because MU_H rises at any particular level of hamburger consumption. Fred buys even more hamburgers, sliding down the higher marginal utility schedule, thereby reducing the marginal utility of hamburgers MU_H , until MU_H / P_H again equals MU_C / P_C . Thus for normal goods the income effect reinforces the substitution effect. Demand curves must slope down.

Suppose hamburgers are an inferior good. Figure 5.A2 then shows a downward shift in the MU_H schedule when the purchasing power of Fred's income increases. At his original consumption bundle, MU_H may fall by more than the fall in P_H , the price of hamburgers. If so, Fred has to *reduce* his hamburger consumption to increase its marginal utility and restore MU_H / P_H to equality with MU_C / P_C as utility maximization requires.

For inferior goods, the income effect goes in the opposite direction to the substitution effect. If the income effect is big enough, it could win out. Lower hamburger prices then reduce the quantity of hamburgers demanded. Demand curves slope upwards. As we discuss in the text, such Giffen goods are rare. It is safe to assume that demand curves slope down in practice.

Modern economists are pretty sniffy about measurable utility, preferring the more general indifference curve analysis used in the text. But indifference curves are tricky the first time you meet them. You need to practise using them to become comfortable with them. Measurable utility, and the simple idea of diminishing marginal utility, allows an easier introduction to the basic properties of demand curves and consumer choice developed in this chapter.

Ordinal utility and indifference curves

If we can give numbers to the utility a consumer receives from consuming a given bundle of goods, then we can write down a mathematical expression between utility measures and every possible bundle of goods faced by our consumer. We call such a mathematical expression a *utility function*. Suppose that the utility function of Fred is $U = xy$, where x denotes the amount of good x and y is the amount of good y . By knowing that function, we know that if Fred is consuming 3 units of good x and 10 units of good y he receives a utility of $U = (3)(10) = 30$ utils from that consumption bundle. And we can do the same calculation for every possible consumption bundle Fred may face. As we know, utility is normally an *ordinal* measure so, while here we can assign numbers to the utility level, no real meaning should be attached to them. The utility function is a way to summarize the information about a consumer's preferences using just a mathematical expression (and this is something economists really like). Why is that? Let's be more general for the moment and write the utility function of Fred as $U(x, y)$.

Consider two different bundles (x_1, y_1) and (x_2, y_2) both containing different units of the two goods x and y . Suppose we know that Fred prefers (x_2, y_2) to (x_1, y_1) . Then, if our utility represents the preferences of Fred, it must be true that:

$$U(x_2, y_2) > U(x_1, y_1)$$

The utility from consuming (x_2, y_2) must be larger than the one obtained from consuming (x_1, y_1) . Suppose instead that we know that Fred is indifferent between consuming (x_1, y_1) and (x_2, y_2) , then it must be true that $U(x_2, y_2) = U(x_1, y_1)$, the two bundles must provide the same level of utility to Fred.

Any mathematical expression that satisfies those basic properties can, then, be used as a utility function to represent a consumer's preferences. In the text we used indifference curves to summarize a consumer's preferences. Here, we see how to derive mathematical expressions for the indifference curves starting from a utility function. An indifference curve shows the combination of consumption bundles that provides the same utility to our consumer. If we know the utility function of our consumer, then it is easy to find his indifference curves. Suppose the utility function is $U = xy$ and suppose we ask: what is the combination of bundles that gives the consumer a utility level of 10?

The answer is given by the combinations that satisfy the following equation:

$$xy = 10.$$

That equation can be written as: $y = \frac{10}{x}$

For example, when $x=10$, the amount of y that gives to our consumer a utility of 10 is simply 1. Therefore, the bundle $(x = 10, y = 1)$ gives a utility level of 10 to our consumer.

There are many other bundles that can give the same level of utility to our consumer; for example, the bundle $(x = 5, y = 2)$.

According to the utility function of our consumer, the bundles $(x=10, y = 1)$ and $(x = 5, y = 2)$ are equally desirable, since they provide the consumer with the same level of utility.

This means that those bundles must lie on the same indifference curve.

Indeed the equation $y = \frac{10}{x}$ is the equation of a possible indifference curve of our consumer. If you plot that equation, you will see that it looks like the typical indifference curve we discussed at the beginning of the chapter. Of course, we can do the same exercise by asking what combination of bundles gives the consumer a utility level of 15 or 22.5. In that way, we can find many indifference curves, each one for a different utility level – as we saw in Figure 5.3.

In general, if the utility function is $U(x, y)$, an indifference curve is defined as the equation $U(x, y) = c$, where c is any positive number (in general, we do not want negative utility!). By changing c , we imply a shift up or down the indifference curve.¹²

A utility function is just a mathematical tool. Nobody really knows her or his utility function from consuming hamburgers and books or any other good. Nevertheless, it is a useful tool that economists use to think about how consumers behave.

MATHS A5.1

UTILITY FUNCTION, MARGINAL UTILITY AND THE MARGINAL RATE

OF SUBSTITUTION

Suppose that the utility function of a consumer is $U = \sqrt{xy}$. Then, we can calculate the marginal utility that our consumer receives from each good he consumes. Suppose our consumer starts with a bundle ($x = 5, y = 5$), which gives him a utility of 5. Now suppose he increases the consumption of x by one unit, from 5 to 6. What is the change in his utility? The utility from the new bundle ($x = 6, y = 5$) is $U = \sqrt{30}$, which is approximately equal to 5.5. So the change in utility is $\Delta U = 5.5 - 5 = 0.5$. This is the marginal utility of consuming an extra unit of good x when we move from bundle ($x = 5, y = 5$) to bundle ($x = 6, y = 5$).

We can continue the process and find the marginal utility of good x by moving from ($x = 6, y = 5$) to ($x = 7, y = 5$), which will be a different number than before, and so on. In calculating the marginal utility of a good, we keep *fixed* the amount of the other good. The marginal utility is the *slope* of the utility function as we hold fixed the quantity of the other good. In general, once we know the utility function we can use calculus to find the expression for the marginal utility. The marginal utility of x can be written as:

$$MU_x = \frac{\partial U}{\partial x}$$

which means that the marginal utility of x (MU_x) is the partial derivative of the utility function U with respect to x . The partial derivative gives us the change in U (∂U) when we change x by a small amount (close to zero) ∂x keeping fixed the amount of y . This is not exactly the definition of marginal utility we use. A small change close to zero is not the same as a change of 1 unit. However, we can consider the equation above as an ‘approximation’ of the true definition of marginal utility. In economics, all the marginal functions are normally calculated using derivatives.

From our utility function we can calculate the marginal utility of x which is equal to: $MU_x = \frac{\partial U}{\partial x} = \frac{1}{2} \sqrt{\frac{y}{x}}$.

With this equation we can calculate the marginal utility at any particular bundle. For example, at bundle ($x=5, y=5$) the marginal

utility of x is $MU_x = \frac{1}{2} \sqrt{\frac{5}{5}}$.¹¹ The same can be done for the marginal utility of good y .

Now suppose that our consumer decides to move from one bundle to another along an indifference curve. This is shown in the figure below.

Suppose he moves from bundle a to bundle b . In doing so, he can consume an extra amount $\Delta x = x_b - x_a$ of good x but he reduces his consumption of good y by $\Delta y = y_b - y_a$. What is the change in his utility in moving from a to b ? Zero. This is because on an indifference curve the level of utility is constant. Let's call MU_x the marginal utility of good x and MU_y the marginal utility of good y . From consuming the extra Δx , utility increases by an amount $MU_x \Delta x$. On the other hand, by reducing consumption of y , utility decreases by an amount $MU_y \Delta y$. Along an indifference curve, these two effects must sum to zero. Along an indifference curve, we have $MU_x \Delta x + MU_y \Delta y = 0$. We can re-write that equation as:

$$\frac{\Delta y}{\Delta x} = -\frac{MU_x}{MU_y} \quad (\text{A2})$$

The left-hand side is the marginal rate of substitution. If the change in x is sufficiently small, then we can write the marginal rate of substitution as the derivative dy/dx , which is the slope of the indifference curve at a particular point. Equation (A.2) says that the marginal rate of substitution is the negative of the ratio of marginal utilities; for example, if the ratio of marginal utilities is 5, the MRS is -5 .

¹¹ A formal proof of this statement is provided at the end of this Appendix.

¹² In mathematical terms what we call an indifference curve is known as the level curve of a function.

CHAPTER 6

Introducing supply decisions

Learning outcomes

By the end of this chapter, you should be able to:

- 1 recognize the legal forms in which businesses are owned and run
- 2 describe revenue, cost, profit and cash flow
- 3 analyse accounts for flows and for stocks
- 4 recognize economic and accounting definitions of cost
- 5 understand whether a firm chooses output to maximize profits
- 6 describe how this choice reflects marginal cost and marginal revenue

Having analysed demand, we turn now to supply. How do firms decide how much to produce and offer for sale? Can a single theory of supply describe the behaviour of different producers, from giant companies such as Microsoft to the self-employed ice cream vendor with a van?

For each possible output level a firm needs to calculate what it *costs* to make this output and how much *revenue* is earned by selling it. At each output, production costs depend on technology, which in turn determines the inputs needed, and on the input prices that the firm faces. Sales revenue depends on the demand curve faced by the firm. The demand curve determines the price for which any output quantity can be sold and thus the revenue the firm earns.

Profits are the excess of revenues over costs. The key to the theory of supply is the assumption that all firms are rational and so they aim to make as much profit as possible. By examining how revenues and costs

change with the level of output produced and sold, the firm chooses the output that maximizes its profits. To understand supply decisions, we then need to analyse what determines revenues and costs.

The assumption of profit maximization is the cornerstone of the theory of supply. We conclude by discussing its plausibility and examine alternative views of what firms' aims might be.

6.1 Business organization

Businesses are self-employed sole traders, partnerships or companies. Sole traders, the commonest type of business organization, operate on a small scale. Partnerships are larger scale; companies are larger still.

A **sole trader** gets the revenue of the business and is responsible for any losses it makes. If he cannot meet these losses, he becomes personally bankrupt. His remaining assets, such as his house, are sold and the money shared out among the creditors.

A **sole trader** is a business owned by a single individual.

A **partnership** is a business jointly owned by two or more people, sharing the profits and being jointly responsible for any losses.

If the business prospers, a sole trader may need money to expand. One way is to bring in new partners, who inject money in exchange for a share of the subsequent profits. **Partnerships** usually have *unlimited liability*. Like sole traders, partners are personally liable for the firm's losses, however large. Firms where trust is involved – solicitors or accountants – are often partnerships. Customers see that the people running the business are willing to put their own wealth behind the firm's obligations.

Any business needs money to start it up and finance its growth. Firms of lawyers, doctors or accountants, businesses relying on human expertise, need relatively little money for such purposes. The necessary funds can be raised from the partners and, possibly, by a bank loan. Businesses requiring large initial expenditure on machinery need much larger initial funds. It is too complicated to have a huge number of partners. Instead, it makes sense to form a company.

Unlike a partnership, a **company** has a legal existence distinct from that of its owners. Ownership is divided among shareholders. The original shareholders may now have sold shares of the profits to outsiders. By selling entitlements to share in the profits, the business can raise new funds.

Shareholders earn a return in two ways. First, the company makes regular *dividend* payments, paying out to shareholders that part of the profits that the firm does not wish to reinvest in the business. Second, the shareholders may make *capital gains* (or losses). If you buy Microsoft shares for a value of £1000 but then people decide Microsoft profits and dividends will be unexpectedly high, the Microsoft shares will increase their market value and you may be able to resell the shares for £1200, for example, making a capital gain of £200. Unlike sole traders and partners, shareholders cannot be forced to sell their personal possessions if the business goes bust – they have **limited liability**. At worst, the shares become worthless.

Companies are run by boards of directors who submit an annual report to the shareholders, who can vote to sack the directors if it seems that other directors could do better. Companies are the main form of organization of big businesses.

A **company** is an organization legally allowed to produce and trade.

Shareholders of a company have **limited liability**. The most they can lose is the money they spent buying shares.

6.2 A firm's accounts

Firms report two sets of accounts, one for **Stocks** and one for **flows**.

Stocks are measured at a given point in time; **flows** are corresponding measures during a period of time.

The water *flowing* out of a tap is different per second and per minute. The measure needs a time interval to make sense of it. The *stock* of water in the basin at any instant is a number of litres, with no time dimension. A firm reports profit-and-loss accounts per year (*flow accounts*) and a

balance sheet showing assets and liabilities at a point in time (*stock* accounts). The two are related, as they are for the basin of water. The inflow from the tap changes the stock of water over time, even though the stock is in litres at each point in time. We begin with flow accounts.

Flow accounts

These ideas are simple, but the calculation of **revenue**, **cost** and **profit** for a large firm is tricky – hence the need for so many accountants. Here is a simple example.

Revenue is what the firm earns from selling goods or services in a given period, **cost** is the expense incurred in production in that period and **profit** is revenue minus cost.

Rent-a-Person (R-a-P) is a firm that hires people whom it then rents out to other firms that need temporary workers. R-a-P charges £10 an hour per worker but pays its workers only £7 an hour. During 2012 it rented 100 000 hours of labour. Business expenses, including leasing an office, buying advertising space and paying telephone bills, were £200000. Table 6.1 shows the *income statement* or *profit-and-loss account* for 2012. Profits before tax were £100000. Tax was £25000. R-a-P's after-tax profits were £75000. Now for the complications.

Unpaid bills

People do not always pay bills on time. At the end of 2012, R-a-P has unpaid bills for workers hired to other firms during the year. Nor has it yet paid its own telephone bill for December. From an economic viewpoint, the right definition of revenues and costs relates to the activities during the year whether or not payments have yet been made.

Actual receipts and payments thus may differ from economic revenue and cost. Profitable firms may still have a poor **cash flow**, for example when customers are slow to pay.

Table 6.1 R-a-P income statement, year to 31 December 2012

Revenue		
100000 hours at £10		£1000000

Revenue		
Cost		
Wages	£700000	
Adverts	£50000	
Office rent	£50000	
Other expenses	£100000	
		–£900 000
Pre-tax profit		£100000
Tax		£25 000
Post-tax profit		£75000

A firm's **cash flow** is the net amount of money actually received during the period.

Physical capital is machinery, equipment and buildings used in production.

Depreciation is the loss in value of a capital good during the period.

Capital and depreciation

R-a-P owns little **physical capital**. Instead, it leases office space, computers and desks. However, many firms do buy physical capital. Economists use 'capital' to denote goods not entirely used up in the production process during the period. Buildings and lorries are capital, to be used again in the next year. Electricity is not capital: purchases in 2009 do not survive into 2010. Economists also use 'durable goods' or 'physical assets' to describe capital goods.

How is the cost of a capital good treated in calculating profit and cost? It is the cost of *using* rather than *buying* capital equipment that is part of the firm's costs within the year.

Suppose R-a-P buys 8 computers in January for £1000 each. £8000 is not the cost of computers in calculating costs and profits for that year. Rather, the cost is the fall in value of the computers over the year. Suppose wear-and-tear and obsolescence reduce the value of a computer

by £300 during the year. Part of the economic cost using 8 computers over the year is the £2400 by which they depreciate during the year.

Depreciation makes economic profit and cash flow differ. When a capital good is first bought there is a large cash outflow, much larger than the depreciation cost of using the good in the first year. Profits may be high but cash flow low. In later years, the firm makes no further cash outlay, having already paid for the capital goods, but must still calculate depreciation as an economic cost since the resale value of goods falls steadily. Cash flow is now higher than economic profit.

Treating depreciation, not the purchase price, as the true economic cost spreads the initial cost over the life of the capital goods but that is not why we calculate cost in this way. R-a-P could have sold its computers for £5600 after a year, restricting its costs to £2400. Since it chose to keep them for reuse in the next year, the latter strategy is even more profitable. Hence the true economic cost of using the computers in the first year is at most £2400.

Inventories

If production is instantaneous, firms can produce to meet orders as they arise. In fact, production takes time. Firms hold **inventories** to meet future demand.

Inventories are goods held in stock by the firm for future sales.

Suppose at the start of 2012 Ford has a stock of 50 000 cars completed and available for sale. In 2012 it makes 1 million new cars and sells 950 000. By December its stock of finished cars is 100 000. What about profit? Revenue arises from selling 950 000 cars. Should cost reflect sales of 950000 cars or the 1 million actually made?

Economic costs relate to the 950 000 cars actually sold. The 50000 cars added to stock are capital the firm made for itself, available for sale in the next period. There was a cash outflow to pay for the manufacture of 1 million cars but part of this cash outflow was used to buy inventories that will provide cash revenue the following year without any new cash outlay on production.

Borrowing

Firms usually borrow to finance their set-up and expansion costs, buying capital goods, solicitors' fees for the paperwork in registering the company, and so on. There is interest to be paid on the money borrowed. This interest is part of the cost of doing business and should be counted as part of the costs.

Stock accounts: the balance sheet

The income statement in Table 6.1 shows flows *in a given year*. We can also examine the firm at *a point in time*, the result of all its past trading operations. The *balance sheet* lists the assets the firm owns and the liabilities for which it is responsible at a point in time. Table 6.2 shows the balance sheet for Snark International on 31 December 2012.

Table 6.2 Snark's balance sheet at 31 December 2012

Assets	£000s	Liabilities	£000s
Cash	40	Accounts payable	90
Accounts receivable	70	Mortgage	150
Inventories	100	Bank loan	60
Factory (bought for 500)	330	Total liabilities	300
		Net worth	240
Total	540		540

Snark's assets are cash in the bank, money owed by its customers (accounts receivable) and inventories in its warehouses and its factory (original cost £500 000, now worth only £330 000 because of depreciation). The total value of Snark's assets is £540 000.

Snark's liabilities are bills it has yet to pay, the mortgage on its factory and a bank loan for short-term cash needs. Its total liabilities (debts) are £300 000. The **net worth** of Snark International is £240 000, its assets minus its liabilities.

A firm's **net worth** is the assets it owns minus the liabilities it owes.

You make a takeover bid for Snark. Should you bid £240 000; that is, its net worth? Probably more. Snark is a live company with good prospects and a proven record. You get not merely its physical and financial assets minus liabilities but also its reputation, customer loyalty and a host of intangibles that economists call *goodwill*. If Snark is a sound company, you should bid more than £240 000. Alternatively, you may think that Snark's accountants undervalued the resale value of its assets. If you can buy Snark for £240 000, you may make a profit by selling off the separate pieces of capital, a practice known as 'asset-stripping'.

CASE 6.1

THE ANATOMY OF A CRISIS: READING THE BALANCE SHEET OF NORTHERN ROCK

Looking carefully at companies' balance sheets is a job for accountants and not for economists. Nevertheless, we can obtain some interesting economic insights by looking at the balance sheet of particular companies. Consider the case of Northern Rock. Northern Rock became famous in September 2007 as the first bank in the UK to be hit heavily by the credit crunch. Northern Rock specialized in property finance activities, especially residential mortgages.

The table above presents a simplified version (where some of the categories have been combined and figures rounded to the nearest billion pounds) of Northern Rock's balance sheet at the end of two different years, 2006 and 2007.

Without going into detail, we can say that between 2006 and 2007 Northern Rock issued about 12 billion loans and advances to customers (mortgages). How could Northern Rock provide these loans to borrowers? By borrowing from other institutions and from the market in the following forms: by using customer accounts, by issuing mortgage-backed securities (these are just Northern Rock's mortgages pooled together and sold to other investors) and by other securities. However, in 2007, Northern Rock was not able to obtain the resources necessary to finance the loans it made. The credit crunch hit the US in the summer of 2007 and spread rapidly to other economies. Raising money became more difficult for

Northern Rock and, on 12 September 2007, it asked for an emergency loan from the Bank of England (the central bank of the UK) for £28 million. This was the signal that things were going pretty poorly for Northern Rock, and indeed, on 14 September 2007, we witnessed the first bank run in the UK for over 100 years, with bank customers queuing outside branches waiting to close their accounts. Northern Rock was nationalized on 22 February 2008. In June 2011 it was put up for sale. On 1 January 2012 it was sold to Virgin Money for £747 million upfront plus £280 million to be paid over the next few years.

		2006	2007
<i>Assets</i>	Loans and advances to banks	6	1
	Loans and advances to customers	87	99
	Investment securities	6	6
	Other assets	2	3
	Total assets	101	109
<i>Liabilities</i>	Loans from central bank	0	28
	Customer accounts	27	12
	Mortgage-backed securities	40	43
	Other securities	24	19
	Other liabilities	7	4
	Equity	3	3
	Total liabilities	101	109



© Cate Gillon | Getty Images

Source: Northern Rock plc Annual Report and Accounts, 2007.

Earnings

When a firm makes profits after tax, it can pay them out to shareholders as dividends or keep them in the firm as retained earnings. **Retained earnings** affect the balance sheet. If kept as cash or used to purchase new equipment, they increase assets. Alternatively, they may reduce the firm's liabilities, by repaying the bank loan. Either way, the firm's net worth increases.

Retained earnings are the part of after-tax profits ploughed back into the business.

Opportunity costs and accounting costs

The income statement and the balance sheet of a company provide two useful guides to how a firm is doing. But economists and accountants take different views of cost and profit. An accountant is interested in tracking the actual receipts and payments of a company. An economist is interested in how revenue and cost affect the firm's supply decision; that is, the allocation of resources to particular activities. Accounting methods can mislead in two ways.

Economists identify the cost of using a resource not as the payment actually made but as its *opportunity cost*. To show that this is the right measure of costs, given the questions economists study, we provide two examples.

If you run your own firm you should take into account the cost of your labour time in the firm. You might draw up an income statement, like the one shown in Table 6.1, find that profits are £20 000 a year and conclude that the firm is a good thing. This conclusion neglects the opportunity cost of your time. If you could have earned £25000 a year working for someone else, being self-employed is losing you £5000 a year despite an accounting profit of £20 000. To understand the incentives that the market provides to guide people towards particular jobs, we must use the economic concept of opportunity cost, not the accounting concept of actual payments.

The second place where opportunity cost must be counted is with respect to capital. You put up the money to start the business. Accounting profits ignore the use of owned (as opposed to borrowed) financial capital. But

this money could have been deposited in an interest-bearing bank account or used to buy shares in other firms. The opportunity cost of that money is part of the *economic* costs of the business but not its accounting costs. If it could earn 5 per cent elsewhere, the opportunity cost of your funds is 5 per cent multiplied by the money you put in. If, after deducting this cost and the true cost of your time, the business still makes a profit, economists call this **supernormal profit**.

The **Supernormal profit** is pure economic profit; it measures all economic costs properly.

Supernormal profits are the true indicator of how well you are doing by tying up your time and funds in the business. Supernormal profits (or losses), not accounting profits (or losses), are the incentive to shift resources into (or out of) a business.

CONCEPT 6.1

ECONOMIC VS ACCOUNTING PROFITS

The inclusion of opportunity costs in economic profits creates an important distinction from the concept of accounting profits. To stress this distinction further, suppose you start your own firm. Suppose that your total revenues are £60 000 and you have explicit costs of £40 000 (for example, wage payments to your workers, the cost of raw materials and so on). According to those numbers, you should obtain an accounting profit of £20 000.

However, suppose that your best alternative was to work for someone else and receive a wage of £25 000. Then your firm, according to an economist, is running at a loss of £5000.

The £25 000 you could have earned somewhere else represents the opportunity cost of your time working in your firm and should be included in the total costs. This opportunity cost enters the economic profits but not the accounting profits.

According to accounting profits, your firm is profitable. According to economic profits, your firm is not profitable. So, in our definition of economic profits we also include in the total costs the remuneration that the owner of the firm obtains by running the firm.

This remuneration is called *normal profit* and it is included in the total cost of our economic profit definition. This is very important because, in many cases, we will say that firms earn zero profits. Zero profit for us will mean zero economic profits. It means that remuneration of the owner is exactly equal to the opportunity cost of running the firm. In our example, suppose that the total revenues were £65 000. The opportunity cost is still £25 000 and the explicit costs are £40 000. In this case, the economic profits are zero.

This does not mean that the owner of the firm gets nothing from his business. He will get a positive remuneration (£25000) but that remuneration is exactly equal to remuneration he could have obtained from his best alternative.

6.3

Firms and profit maximization

Economists assume that firms choose how much to produce in order to *maximize profits*. Some economists and business executives question this assumption. For example, a sole owner may prefer to work for himself even if he could earn more in total by working somewhere else. His business decisions reflect maximization of his total job satisfaction not merely his monetary profit.

Ownership and control

A more significant reason to question profit maximization comes from considering the case of large firms. A large firm is run not by its owners but by a salaried board of directors. The directors are the experts with the relevant information on whether the firm is well managed or not. At the annual meeting, shareholders may dismiss the board; doing so is rare, however.

Economists call this a separation of ownership and control. Although shareholders want the maximum possible profit from the firm's activity, the directors who actually make the decisions can pursue different objectives. Do directors have an incentive to act other than in the interests of the shareholders?

Directors' salaries are usually higher, the larger the firm. Directors may aim for size and growth rather than the maximum possible profit, spending large sums on costly advertisements to boost sales.

The A principal or owner may delegate decisions to an agent. If it is costly for the principal to monitor the agent, the agent has inside information about its own performance, causing **a principal agent problem**.

The separation of ownership and control in companies leads to what economists call a **principal agent problem**. The agents (here, the directors) are tempted to act in their own interests rather than those of their principals (the shareholders).

Nevertheless, there are two reasons why the aim of profit maximization is a good place to start, including for large firms. Even if the shareholders cannot recognize that profits are lower than they might be, other firms with experience in the industry may catch on faster. If profits are low, share prices will be low. By mounting a takeover, another company can buy the shares cheaply, sack the existing managers, restore profit-maximizing policies and make a handsome capital gain as the share price rises once the stock market sees the improvement in profits. Fear of takeover may induce directors to try to maximize profits.

Moreover, aware of the scope for directors' discretion, shareholders try to ensure that the interests of directors and shareholders coincide. By giving senior directors big bonuses tied to profitability or share performance – a small cost when spread over many shareholders but a major incentive for the existing management – shareholders try to make senior management care about profits as much as shareholders do.

The assumption that firms try to maximize profits is more robust than might first be imagined. Before using it to develop the theory of supply, we discuss the stock market in more detail.

CONCEPT 6.2

HOSTILE TAKEOVERS

A hostile takeover is an acquisition in which the firm being purchased does not want to be purchased, or does not want to be

purchased by the particular buyer making the bid. How is it possible to buy something that is not for sale? Hostile takeovers only work with publicly traded firms. Those are firms that have issued stock that can be bought and sold on public stock markets. The stock of a firm is divided into shares. If a firm has issued 100 shares and you buy 51 of them, then you own a majority and in many respects you now control that firm. This is a possible way for a hostile takeover to take place. The buyer can gain control by acquiring in the market the majority of shares from the existing shareholders of the target firm.

In Germany, hostile takeovers have traditionally been rare. In contrast, many UK takeovers are hostile bids uninvited by existing managers. Some economists see hostile bids as a vital force for efficiency. The threat of hostile takeovers deters managers from departing too far from the profit-maximizing policies that shareholders want. Slack management leads to low profits, depressed share prices and opportunities for takeover raiders to buy the company cheaply. The threat of takeover provides a discipline that helps overcome the principal–agent problem.

Obviously not all hostile takeovers end successfully, since the existing shareholders may refuse the offer made by the buyer. An example of an unsuccessful takeover was the attempt by Microsoft to gain control of Yahoo! On 1 February 2008, Microsoft made an unsolicited bid to purchase Yahoo! This offer was rejected on 10 February and on 3 May Microsoft finally withdrew the offer.

6.4 The firm's supply decision

Firms produce goods and services that are sold in markets. We want to understand how a firm decides how much to produce of a given good or service.

Suppose a firm makes spoons. The firm needs to decide how many spoons to produce and sell. The first thing that the firm should consider is how costly it is to produce the spoons. Some ways to make spoons use lots of labour and few machines; other ways use many machines but little labour. The firm knows different techniques for making spoons and the

cost of hiring inputs – the wage rate for workers and rental for leasing a machine. The second thing that the firm must consider is the demand condition. The firm knows its demand curve. This is the demand curve derived from all the customers who want to buy the spoons made by that particular firm. If the firm knows the demand curve it faces, then it knows its revenue from selling different quantities of spoons at different prices.

Knowing costs and revenues generated by different amounts of spoons produced, the firm is able to find the profit generated by those amounts, since profit is simply revenues minus costs.

The firm chooses the level of output (here, the number of spoons produced and then sold in the market) in order to maximize its profits. Changing the level of output produced affects both the costs of production and the revenues from sales. Costs and demand conditions jointly determine the output choice of a profit-maximizing firm.

Cost minimization

Closely related to the idea of profit maximization is the concept of cost minimization. Indeed, profit maximization and cost minimization can be seen as two equivalent concepts. A profit-maximizing firm certainly wants to make its chosen output level at the minimum cost possible. By producing the same output at lower cost, it could increase profits. Thus a profit-maximizing firm must produce its chosen output as cheaply as possible.¹

Total cost

Knowing the available production methods and the costs of hiring workers and machines, the firm calculates the least cost at which each output can be made. It is not worth using many machines to make only a few spoons; to make more spoons, it makes sense to use more machines.

Table 6.3 shows various outputs in column (1). Column (2) shows the minimum cost at which each output can be made. The firm incurs a cost of £10 even when output is zero. This is the cost of being in business at all – running an office, renting a telephone line and so on. Thereafter, costs rise with output. Costs include the opportunity costs of all resources used in production. Total cost is higher, the more is produced.

At high levels of output, cost rises sharply as output increases: the firm has to pay the workers overtime to work weekends and nights.

Total revenue

The total revenue the firm obtains from an output depends on price and hence demand. Column (3) of Table 6.3 summarizes the demand curve faced by the firm; it shows the price at which each output can be sold. Column (4) calculates sales revenue (or total revenue); that is, price multiplied by quantity. At a price of £21 the firm sells only one spoon, while at a price of £12 the firm is able to sell 10 spoons. The lower the price, the greater the sales: its demand curve slopes down.

Table 6.3 Cost, revenue, profit (weekly)

(1) Output	(2)Total cost (£)	(3) Price (£)	(4)Total revenue (1) x (3) (£)	(5) Profit (4) – (2) (£)
0	10	–	0	– 10
1	25	21	21	–4
2	36	20	40	4
3	44	19	57	13
4	51	18	72	21
5	59	17	85	26
6	69	16	96	27
7	81	15	105	24
8	95	14	112	17
9	111	13	117	6
10	129	12	120	–9

Profit

Column (5) of Table 6.3 shows profit, the difference between total revenue and total cost. At low output, profit is negative. At the highest output of 10, profit is again negative. At intermediate outputs, the firm makes positive profit.

The highest profit is £27 a week, at an output of 6 spoons. At £16 each, total revenue is £96. Production cost, properly calculated, is £69, leaving

a profit of £27 a week. Therefore we can say that our firm should produce 6 spoons a week, since, at that level of output, profits are the highest possible and thus maximized. This chosen output, or supply decision, is the highlighted row in Table 6.3.

Notice that maximizing profit is not the same as maximizing revenue. By selling 10 spoons a week the firm could earn £120, but it would cost £129 to make them. Making the last few spoons is expensive and brings in little extra revenue. It is more profitable to make fewer.

6.5

Marginal cost and marginal revenue

It is helpful to view the same problem of profit maximization from a different angle. At each output level, we now ask whether the firm should increase output still further. Suppose the firm makes 3 spoons and considers making 4 spoons. Table 6.3 shows this raises total cost from £44 to £51, a £7 increase in total cost. Revenue rises from £57 to £72, a rise of £15. Raising output from 3 to 4 spoons adds more to revenue than to cost. Profit rises by £8 (£15 more revenue minus £7 more cost). The firm then checks if it is also profitable to increase production from 4 to 5, and so on.

This approach – examining how 1 more unit of output affects profit – focuses on the **marginal cost** and **marginal revenue** of producing 1 more unit.

Marginal cost is the rise in total cost when output rises by 1 unit.

Marginal revenue is the rise in total revenue when output rises by 1 unit.

If marginal revenue exceeds marginal cost, the firm should raise output. Producing and selling an extra unit adds more to total revenue than to total cost, raising total profit. If marginal cost exceeds marginal revenue, the extra unit of output reduces total profit.

Thus we can use marginal cost and marginal revenue to calculate the output that maximizes profit. As long as marginal revenue exceeds marginal cost, keep increasing output. As soon as marginal revenue falls short of marginal cost, stop increasing output.

Marginal cost

Table 6.4 uses Table 6.3 to calculate the marginal cost of producing each extra unit of output. Increasing output from 0 to 1 raises total cost from £10 to £25. The marginal cost of the first unit is £15. Increasing production from 1 to 2 spoons raises total cost from £25 to £36, meaning that the marginal cost of the second unit produced is £11. Table 6.4 shows this marginal cost of each output level; that is, the extra total cost of raising output by the last unit.

Table 6.4 Total and marginal cost

Output	Total cost (£)	Marginal cost (£)
0	10	—
1	25	15
2	36	11
3	44	8
4	51	7
5	59	8
6	69	10
7	81	12
8	95	14
9	111	16
10	129	18

Marginal cost is large when output is low, but also when output is high. Marginal cost is lowest when making the fourth unit, which adds only £7 to total costs.

As output increases, why do marginal costs start high, then fall, then rise again? The answer reflects different production techniques. At low output, the firm uses simple techniques. As output rises, more sophisticated machines are used, making extra output quite cheap. As output rises still further, the difficulties of managing a large firm emerge. Raising output gets hard and marginal costs rise.

Figure 6.1 plots this relation between output and marginal cost. The marginal cost curve can be different from firm to firm. In a coal mine that is nearly worked out, marginal cost rises steeply with extra output.

In mass-production industries, as output increases marginal cost may decline and then become constant (see Figure 6.1 again).

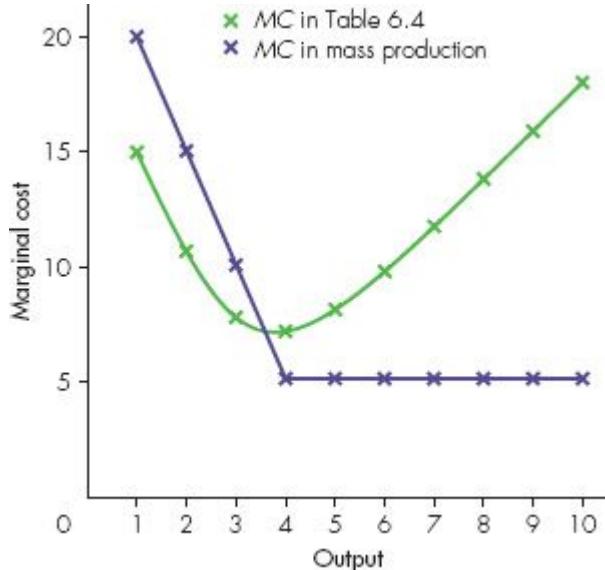


Figure 6.1 Marginal cost curves

CASE 6.2

MARGINAL COSTS IN PRACTICE: THE CASE OF LOCAL BUS TRANSPORTATION IN THE UK

In reality, calculating marginal costs for a firm is not that straightforward. To do so, we need to have an idea of what the firm's total costs look like and of how those total costs are related to the output produced. In practice, we need to estimate how total costs change as output is increased or decreased.

Concessionary passengers are those who pay a reduced ticket price or no price at all for public transportation. What is the marginal cost of having an extra concessionary passenger on a local bus? There are three main categories of marginal cost:

- Fuel, tyres and oil:* an additional passenger increases the weight of the bus, which will lead to an increase in fuel and oil

consumption and greater wear of tyres.

- (a) *Maintenance and cleaning*: an extra passenger is likely to increase wear and tear of the bus interior.
- (a) *Time*: an extra passenger increases boarding and alighting time and acceleration and deceleration time associated with additional stops.

Other categories are related to insurance and the provision of information about the concessionary scheme.

The UK Department of Transport estimates those marginal costs in order to calculate the amount of reimbursement that local bus companies are entitled to receive for carrying concessionary passengers.* The estimates for 2012/13 are shown below.

Marginal cost component	Estimated value in pence (2009/10 prices)
Fuel, tyres and oil	0.4
Maintenance and cleaning	0.1
Time	1.3
Insurance	2.7
Information	0.5
Total	5.0

As a result, the marginal cost of a concessionary passenger taking a local bus is estimated to be £5.

* This is an example of regulation and will be discussed in more detail in Chapter 8.

Marginal revenue

Still based on Table 6.3, Table 6.5 shows marginal revenue; that is, the extra total revenue when an extra unit of output is made and sold. Raising output from 0 to 1 raises revenue from £0 to £21. The marginal revenue of the first unit is £21. Raising output from 7 to 8 units raises revenue from £105 to £112, so marginal revenue is £7. Total revenue and marginal revenue depend on the demand curve for the firm's product.

Marginal revenue, also shown in Figure 6.2, falls steadily as output rises and can be negative at high output levels. To sell 11 spoons, the price must be cut to £10 each. Total revenue is £110. Since 10 spoons earn £120 in Table 6.5, the marginal revenue from moving from 10 to 11 spoons is £110 – £120, that is, –£10.

Marginal revenue = extra revenue from making and selling 1 more unit of output

Table 6.5 Price, total revenue and marginal revenue

Output	Price (£)	Total revenue (£)	Marginal revenue (£)
0	–	0	–
1	21	21	21
2	20	40	19
3	19	57	17
4	18	72	15
5	17	85	13
6	16	96	11
7	15	105	9
8	14	112	7
9	13	117	5
10	12	120	3

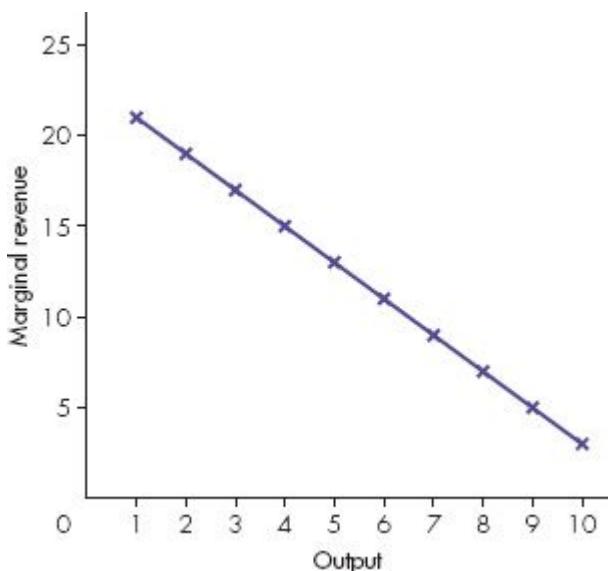


Figure 6.2 Marginal revenue

Marginal revenue is the increase in the firm's revenue from an increase in sales by one unit. If the firm can sell more output only by reducing its price, marginal revenue declines as output rises.

The shape of the marginal revenue curve reflects the shape of the firm's demand curve. Demand curves slope down. To sell more output, the price must be cut. Selling an extra unit of output at this lower price is the first component of marginal revenue. However, to sell that extra unit the firm has to cut the price for which *all* previous units of output can be sold. This effect reduces the marginal revenue obtained from selling an extra unit of output.

Marginal revenue falls steadily for two reasons. First, because demand curves slope down, the extra unit must be sold at a lower price. Second, successive price reductions reduce the revenue earned from *existing* units of output, and at larger output there are more existing units on which revenue is lost when prices fall further. To sum up, (a) marginal revenue falls as output rises and (b) marginal revenue is less than the price for which the last unit is sold, because a lower price reduces revenue earned from existing output (see Maths 6.1).

MR, MC and the output choice

Combining marginal cost (MC) and marginal revenue (MR), Table 6.6 examines the output that maximizes the firm's profits. If MR exceeds MC , a 1-unit increase in output will increase profits. The last column shows that this reasoning leads the firm to make at least 6 units of output. The firm now considers increasing output from 6 to 7 units. Marginal revenue is £9 and marginal cost is £12. Profits fall by £3. Output should *not* be expanded to 7 units, or to any level above this.

Table 6.6 Using marginal revenue and marginal cost to choose output

Output	MR (£)	MC (£)	$MR - MC$ (£)	Output decision
1	21	15	6	Raise
2	19	11	8	Raise
3	17	8	9	Raise
4	15	7	8	Raise
5	13	8	5	Raise
6	11	10	1	
7	9	12	-3	Lower
8	7	14	-7	Lower
9	5	16	-11	Lower
10	3	18	-15	Lower

The firm should expand up to 6 units of output but no further. This output maximizes profits, as we know already from Table 6.5.

Table 6.3, based on total cost and total revenue, and Table 6.6, based on marginal cost and marginal revenue, are different ways to study the same problem. Economists frequently use marginal analysis. Is there a small change that could make the firm better off? If so, the current position cannot be the best possible one and changes should be made.

Marginal analysis should be subjected to one very important check. It may miss an all-or-nothing choice. For example, suppose that MR exceeds MC up to an output level of 6 units but thereafter MR is less than MC . Six units is the best positive output level. However, if the firm incurs large costs whether or not it produces (for example, a vastly overpaid managing director), the profit earned from producing 6 units may not cover these fixed costs. Conditional on paying these fixed costs, an output level of 6 units is then the loss-minimizing output level. Shareholders might do better to shut the firm and fire the fat cat boss. We examine this issue in the next chapter.

To sum up, a profit-maximizing firm should expand output so long as marginal revenue exceeds marginal cost but stop expansion as soon as marginal cost exceeds marginal revenue. This rule guides the firm to the best positive level of output. If the firm is not making profits even in this position, it may do better to close down altogether.

MATHS 6.1

TOTAL AND MARGINAL REVENUE WITH A LINEAR DEMAND

Consider a firm that faces a downward-sloping linear inverse demand for its produced good. Suppose the inverse demand function is $P = a - bQ$, where P is the price, Q is the quantity produced by the firm, $a > 0$ is the intercept and $b > 0$ is the slope of the inverse demand.

The total revenue function for the firm is given by

$$TR(Q) = P \times Q \quad (1)$$

where $TR(Q)$ stands for total revenue and the Q in brackets means that the total revenue depends on the quantity produced. As the quantity produced changes, so the total revenue changes. Using the inverse demand to substitute P into the total revenue function gives us:

$$TR(Q) \equiv (a - bQ) \times Q = -bQ^2 + aQ \quad (2)$$

From the total revenue function in (2) we can see that total revenue is zero when the firm does not produce ($Q = 0$) and when the firm produces an amount $Q = a/b$. You should notice that $Q = a/b$ represents the horizontal intercept of the inverse demand. At that quantity the price is going to be zero and so total revenue is zero as well. Between $Q = 0$ and $Q = a/b$ total revenue first increases and then decreases as Q increases.

Once we know the total revenue function, we use calculus to find the marginal revenue function. The marginal revenue tells us by how much the total revenue will change if we increase the quantity by 1 unit.

The marginal revenue function can be found by taking the derivative of the total revenue function with respect to Q :

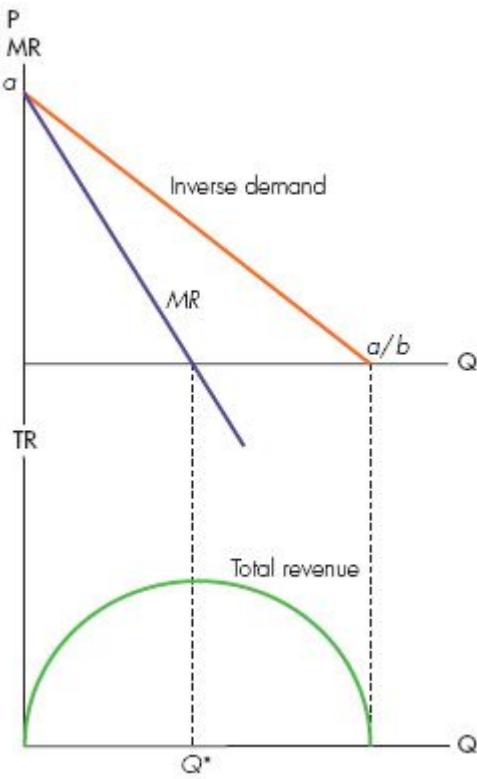
$$MR(R) = \frac{dTR(Q)}{dQ} \quad (3)$$

Equation (3) tells us by how much the total revenue changes ($dTR(Q)$) if we change the quantity produced by a very small amount (close to zero) dQ . This is not exactly the definition of marginal revenue we use. A small change close to zero is not the same as a change of 1 unit. However, we can consider equation (3) as an ‘approximation’ of the true definition of the marginal revenue function. Indeed we normally use derivatives to calculate every marginal function (marginal cost, marginal utility and so on).

The marginal revenue function associated with the total revenue function in (2) is:

$$MR(Q) = a - 2bQ \quad (4)$$

Equation (4) is particularly interesting. First, it is linear in Q . Second, it looks like exactly the inverse demand function but with a slope that is twice as steep. Indeed, this is a general result. When a firm faces a linear inverse demand function, the marginal revenue function of that firm is the inverse demand function with a slope that is twice as steep. In the figure on the right we plot the inverse demand, the total revenue function generated by that demand and the marginal revenue function when the inverse demand is linear.



In the top graph, we plot the inverse linear demand and the marginal revenue function. The marginal revenue function has the same vertical intercept as the inverse demand but the slope is twice as steep and so it is steeper than the inverse demand slope. In the bottom graph we plot the total revenue function. Total revenue is zero when $Q = 0$ and then it increases as Q increases. It reaches a maximum at Q^* and then, as Q increases above Q^* , it decreases and becomes zero when $Q = a/b$. Between $Q = 0$ and Q^* , TR increases as Q increases and so MR is positive.

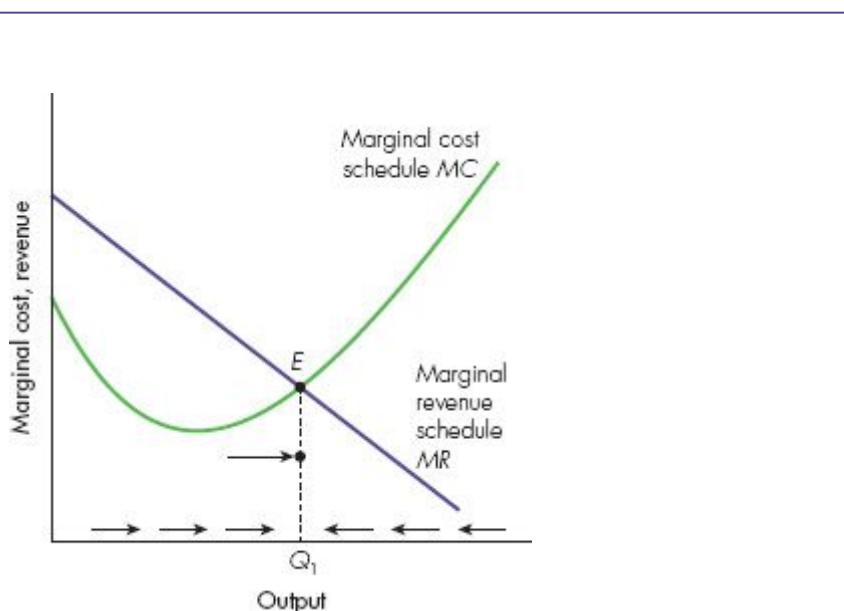
Between Q^* and $Q = a/b$, TR decreases as Q increases and MR becomes negative. For each level of output Q , the marginal revenue is lower than the price and it decreases steadily as Q increases.²

6.6

Marginal cost and marginal revenue curves

Thus far we have assumed the firm produces an integer number of goods, such as 0, 1 or 2, rather than a quantity such as 1.5 or 6.7. Output is not usually confined to integer levels. For goods such as wheat or milk, the firm can sell in odd amounts. Even for goods such as cars, sold in whole units, the firm may be selling 75 cars every four weeks, or 18.75 cars a week. It is convenient to imagine that firms can vary output and sales levels continuously.

We can then draw smooth schedules for marginal cost MC and marginal revenue MR , as shown in Figure 6.3. Profits are maximized where the schedules cross, at point E . The output Q_1 maximizes profits (or minimizes losses). At smaller outputs, MR exceeds MC and expansion increases profits (or reduces losses).



The marginal cost and marginal revenue schedules are shown changing smoothly. The firm's optimal output is Q_1 , at which marginal revenue is equal to marginal cost.

Anywhere to the left of Q_1 , marginal revenue is larger than marginal cost and the firm should increase output, as shown by the arrows. Where output is greater than Q_1 , marginal revenue is less than marginal cost and profits are increased by reducing output. If the firm is losing money at Q_1 it has to check whether it might be better not to produce at all than to produce Q_1 .

Figure 6.3 Marginal cost and marginal revenue

To the right of Q_1 , MC exceeds MR . Expansion adds more to costs than revenue, and contraction saves more in costs than it loses in revenue. The profit incentive to increase output to the left of Q_1 and to reduce output

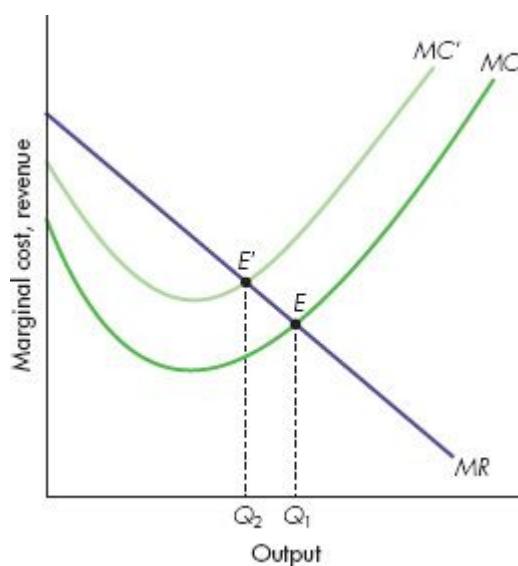
to the right of Q_1 is shown by the arrows in Figure 6.3. This incentive guides the firm to choose Q_1 , provided the firm should be in business at all. At Q_1 , marginal revenue is exactly equal to marginal cost. Table 6.7 summarizes the conditions for determining the output that maximizes profits.

Table 6.7 The firm's output choice

Marginal condition	Output decision	Check
$MR > MC$	Raise	
$MR < MC$	Cut	
$MR = MC$	Stay	If profits > 0 , make this output. If not, quit.

Changes in cost

Suppose the firm faces a price rise for a raw material. At each output, marginal cost is higher than before. Figure 6.4 shows this upward shift from MC to MC' . The firm now produces at E' . Higher marginal costs reduce profit-maximizing output from Q_1 to Q_2 .



The marginal cost curve shifts up from MC to MC' as a result of an increase in the costs of using a factor of production; for instance, the wage may have risen. This upward shift moves the intersection of MC and MR curves from E to E' . Output falls from Q_1 to Q_2 . Thus, when the firm's costs rise, it decides to produce less.

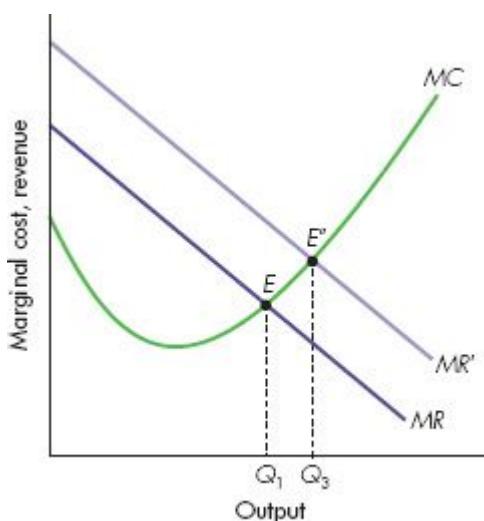
Figure 6.4 An increase in marginal cost reduces output

A demand shift

Suppose the firm's demand curve shifts up, for example because the good produced by the firm becomes more popular and so more consumers want to buy it. If the demand shifts up, the marginal revenue curve must also shift up. At each output, price and marginal revenue are higher than before. In Figure 6.5 the MR curve shifts up to MR' , inducing the firm to move from E to E' . Higher demand makes the firm expand output from Q_1 to Q_3 . Notice that, as the demand increases, so too does the price at which the firm can sell each level of output. Figure 6.4 shows us that a profit-maximizing firm will respond to this increase in the price by increasing the output produced, a result that is consistent with the idea that a supply curve should be positively sloped, as discussed in Chapter 3.

Do firms know their marginal cost and marginal revenue curves?

Do firms in the real world know their marginal cost and marginal revenue curves, let alone go through some sophisticated calculations to make sure output is chosen to equate the two?



When the MR curve shifts upwards from MR to MR' , the intersection point between the MR and MC curves shifts from E to E' . The firm's optimal level of output increases from

Q_1 to Q_3 . The upward shift in the marginal revenue curve could result, for instance, from an increase in the number of customers in the firm's market.

Figure 6.5 An upward shift in marginal revenue increases output

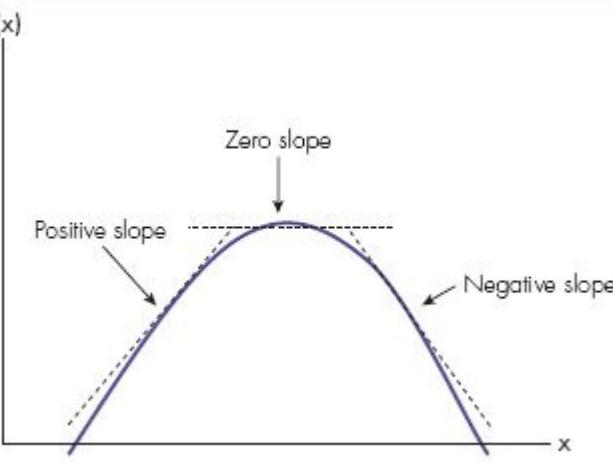
Such thought experiments by firms are not necessary for the relevance of our model of supply. If, by luck, hunch or judgement, a firm succeeds in maximizing profits, marginal cost and marginal revenue *must* be equal. Our formal analysis merely tracks the hunches of smart managers who get things right and survive in a tough business world.

In this chapter we introduced cost and revenue conditions and the idea of profit maximization. Later chapters fill in the details but we now have the basis for a theory of how much output firms choose to supply. Firms choose the level of output that maximizes profits. At this level of output, marginal cost equals marginal revenue.

MATHS 6.2

THE MATHEMATICS OF PROFIT MAXIMIZATION

We have learnt that a firm maximizes its profits when it chooses the level of output at which the marginal revenue is equal to the marginal cost. Here, we show how to derive this result mathematically. First, we need to understand what it means to find the maximum of a function. Finding a maximum is like finding the top of a hill. You first need to climb the hill, meaning that the slope of the hill must be positive. At a certain point you reach the top of the hill (the *maximum*), and at that point you must be on a flat spot, meaning that the slope of the hill at the top must be zero. If you continue walking you must go downhill, meaning that the slope of the hill becomes negative after the top. This simple reasoning gives us an idea of how to find the maximum of a function, as depicted in the figure below. This figure shows the graph of a function $f(x)$, which we depict as hill-shaped.



A profit function will normally look like the function displayed above. To find the maximum of that function, we need to find the point at which the slope of the function is zero. However, this is normally not enough to identify a maximum. Indeed, if instead of a hill you are facing a valley (just turn the graph in the figure above upside down), the bottom of the valley (which we call a *minimum*) is flat and therefore the slope is zero not only at a maximum but also at a minimum. Of course, we can always try to plot our function and see if it looks like a hill (in that case, where the slope is zero there will be a maximum) or a valley (in that case, we find a minimum). But there is no need to plot the function. We can use the fact that around a maximum it must be true that the slope of the function is decreasing (it must first be positive, then zero at the maximum, and then negative). Similarly, around a minimum it must be true that the slope of the function is increasing (it is first negative, becomes zero at the minimum and then becomes positive).

Let's apply this reasoning to the idea of profit maximization. We write the profit function as:

$$\pi(Q) = TR(Q) - TC(Q)$$

where π denotes profits, which are given by the difference between total revenues (TR) and total costs (TC). We write $\pi(Q)$ to denote that profits depend on the quantity produced Q , as total revenues and total costs. The firm faces the problem of finding the quantity to produce Q at which the function $\pi(Q)$ reaches a maximum. We

need to find where the slope of the function $\pi(Q)$ is zero. The slope is just the derivative of $\pi(Q)$ with respect to Q : $\frac{d\pi(Q)}{dQ}$

So we need to find the Q at which $\frac{d\pi(Q)}{dQ} = 0$

This is called the *first-order condition* for a maximum.

Since $\pi(Q) = TR(Q) - TC(Q)$, we have that:

$$\frac{d\pi(Q)}{dQ} = \frac{dTR(Q)}{dQ} - \frac{dTC(Q)}{dQ}$$

We know (from Maths 6.1) that $\frac{dTR(Q)}{dQ}$ is the marginal revenue from producing an amount Q . The term $\frac{dTC(Q)}{dQ}$

tells us by how much total costs change ($dTC(Q)$) if we change the quantity produced by a small amount (dQ). But this is what we call a marginal cost. So we

can define: $\frac{dTR(Q)}{dQ} = MR(Q)$ and $\frac{dTC(Q)}{dQ} = MC(Q)$.

The first-order condition implies:

$$\frac{d\pi(Q)}{dQ} = MR(Q) - MC(Q) = 0 \quad (1)$$

This implies that, at the amount Q where the slope of the profit

function is zero $\left(\frac{d\pi(Q)}{dQ} = 0\right)$, it must be true that $MR(Q) = MC(Q)$; that is, the marginal revenue from producing that amount must be equal to the marginal cost of producing it. Therefore the value of Q that satisfies the first-order condition in equation (1) is the amount that a firm must choose in order to maximize profits.

To see that the solution of equation (1) is indeed a maximum, we need to check that the slope of the profit function is decreasing around the maximum. In practice, when we say that the slope of a function is decreasing around a maximum, we say that the slope of the slope must be negative at the maximum. What is the slope of the slope? It is the derivative of the slope at a given point. The derivative of a derivative is called the second-order derivative. For

the profit function, the second-order derivative is: $\frac{d^2\pi(Q)}{dQ^2}$, where the 2 denotes the fact that this is the second-order derivative. So $\frac{d^2\pi(Q)}{dQ^2}$, is just the derivative with respect to Q of $\frac{d\pi(Q)}{dQ}$.

If the Q that satisfies the first-order condition in equation (1) is a maximum for the profit function $\pi(Q)$, then at that Q it must be true that:

$$\frac{d^2\pi(Q)}{dQ^2} < 0 \quad (2)$$

Equation (2) is called the *second-order condition* for a maximum. (For a minimum, the second-order condition is reversed; we want the second-order derivative of the function to be positive at the minimum point.)

The second-order condition in our case can be written as:

$$\frac{d^2\pi(Q)}{dQ^2} = \frac{d^2TR(Q)}{dQ^2} - \frac{d^2TC(Q)}{dQ^2} < 0$$

which means that the slope of the marginal revenue minus the slope of the marginal cost must be negative at the Q that maximizes profit. Another way to say this is that the marginal cost must intersect the marginal revenue from below at the profit-maximizing output. This is what is demonstrated, for example, in Figure 6.3 and the figures thereafter.

Summary

- The **theory of supply** is the theory of how much output firms choose to produce.
- There are three types of firm: self-employed **sole traders**, **partnerships** and **companies**. Sole traders are the most numerous but are often very small businesses. The large firms are companies.

- Companies are owned by their shareholders but run by the board of directors.
- Shareholders have **limited liability**. Partners and sole traders have **unlimited liability**.
- **Revenue** is what the firm earns from sales. **Costs** are the expenses incurred in producing and selling. **Profits** are the excess of revenue over costs.
- Costs should include opportunity costs of all resources used in production. **Opportunity cost** is the amount an input could obtain in its next-highest-paying use. In particular, economic costs include the cost of the owner's time and effort in running a business. **Economic costs** also include the opportunity cost of financial capital used in the firm. **Supernormal profit** is the pure profit accruing to the owners after allowing for all these costs.
- Firms are assumed to aim to **maximize profits**. Even though the firm is run by its managers, not its owners, profit maximization is a useful assumption in understanding the firm's behaviour. Firms that make losses cannot continue in business indefinitely.
- In aiming to maximize profits, firms necessarily produce each output level as cheaply as possible. Profit maximization requires minimization of costs for each output level.
- Firms choose the **optimal output level** to maximize total economic profits. This decision can be described equivalently by examining marginal cost and marginal revenue. **Marginal cost** is the increase in total cost when one more unit is produced. **Marginal revenue** is the corresponding change in total revenue and depends on the demand curve for the firm's product. **Profits are maximized at the output at which marginal cost equals marginal revenue**. If profits are negative at this output, the firm should close down if doing so reduces losses.
- An upward shift in the marginal cost curve reduces output. An upward shift in the marginal revenue curve increases output.

- It is unnecessary for firms to calculate their marginal cost and marginal revenue curves. Setting MC equal to MR is merely a device that economists use to mimic the hunches of smart firms who correctly judge, by whatever means, the profit-maximizing level of output.

Review questions



EASY

- 1 R-a-P is run by an owner, who can earn £40 000 a year by managing another firm. She has also invested £200 000 in R-a-P that could be earning 12 per cent interest elsewhere. What are the economic profits of R-a-P? (Use the data given in the table below and calculate economic profit based on pre-tax profit.)

R-a-P income statement, year to 31 December 2012

Revenue		
100000 hours at £10		£1000000
Cost		
Wages	£700000	
Adverts	£50000	
Office rent	£50000	
Other expenses	£100000	
		−£900 000
Pre-tax profit		£100000
Tax		£25 000
Post-tax profit		£75000

- 2 (a) Snark International borrows £50 000 from the bank and increases its inventories. How does this affect its balance sheet (shown in the table below)? (b) How would interest on the loan appear in the income statement of Snark International?

Snark's balance sheet at 31 December 2012

Assets	£000s	Liabilities	£000s

Assets	£000s	Liabilities	£000s
Cash	40	Accounts payable	90
Accounts receivable	70	Mortgage	150
Inventories	100	Bank loan	60
Factory (bought for 500)	330	Total liabilities	300
		Net worth	240
Total	540		540

- 3 (a) Do firms aim to maximize profits? (b) Should firms support charities, the arts and political campaigns? Is there any conflict with (a)?
- 4 **True or False** If $MR < MC$ at a given level of output, a firm should increase production to increase profits.
- 5 **Common fallacies** Why are these statements wrong? (a) Firms with an accounting profit must be thriving. (b) Firms do not know their marginal costs. A theory of supply cannot assume that firms set marginal revenue equal to marginal cost. (c) To maximize profit, maximize sales.

MEDIUM

- 6 A firm with the costs shown in the following table can now sell as much output as it wants at a price of £12. (a) Draw the MR and MC curves. (b) What output will it produce?

Output	Total cost (£)	Marginal cost (£)
0	10	
1	25	15
2	36	11
3	44	8
4	51	7
5	59	8
6	69	10
7	81	12
8	95	14
9	111	16

Output	Total cost (£)	Marginal cost (£)
10	129	18

- 7 How do the following affect the income statement for R-a-P shown in the table in Question 1? (a) R-a-P owes £70 000 to its workers for work done in the year. (b) Instead of renting an office (for the rent of £50 000 shown under Cost), R-a-P owns its office. (c) During the year R-a-P was paid by a creditor owing money from the year before.
- 8 In the table below, assume that the total costs of making each unit of output, as shown in the second column, increase by £40. What level of output should the firm produce? Explain.

Cost, revenue, profit (weekly)

[1] Output	[2] Total cost (£)	[3] Price (£)	[4] Total revenue [1] × [3] (£)	[5] Profit [4] – [2] (£)
0	10	—	0	-10
1	25	21	21	-4
2	36	20	40	4
3	44	19	57	13
4	51	18	72	21
5	59	17	85	26
6	69	16	96	27
7	81	15	105	24
8	95	14	112	17
9	111	13	117	6
10	129	12	120	-9

- 9 A firm faces the following linear inverse demand for its product, $P = 60 - 2Q$. Find the firm's total revenue function, $TR(Q)$. Plot the total revenue function. (*Hint:* using $P = 60 - 2Q$, plot a demand schedule for an arbitrary number of quantities – say, from 5 to 25 in fives.)
- 10 The following table reports the total revenue and the total cost of Keinko International, a firm producing coffee. Keinko does not have any fixed costs. Complete the columns for marginal revenue (MR) and marginal cost (MC). In a graph, plot the MR and MC curves and show the profit-maximizing level of output.
-

Quantity	TR	TC	MR	MC
1	48	5	—	—
2	90	20		
3	128	45		
4	163	80		
5	195	125		
6	225	180		
7	251	245		
8	273	320		
9	293	405		
10	309	500		

- |1 Now suppose that Keinko faces an increase in demand for its coffee. For *each unit* of coffee, total revenues increase by 20. Find the new total revenue for each level of output and the corresponding new marginal revenue. In a graph, plot the new *MR* and *MC* curves. How has the increase in demand affected the output choice of Keinko?

HARD

- |2 **Essay question** ‘The industrial revolution was built on the ability of entrepreneurs to float companies and obtain funding. Today, it is often argued that stock exchanges force firms to be focused too much on the short term, making it hard to raise long-term funds. Private equity firms see themselves as addressing this shortcoming of stock markets. The amazing thing about private equity is not its recent appearance but that it took so long to appear.’ Discuss.
- |3 A firm faces the following linear demand for its product, $QD = 30 - P/2$. The firm has a marginal cost of production given by $MC=8$. Find the expression for the firm’s marginal revenue. Plot the *MC* and *MR* curves on a graph. What is the amount of output that the firm should produce? At what price is the output sold?
- |4 A firm’s revenue function is given by $TR(Q) = 10Q$ and its cost function by $TC(Q) = 150 + 2Q^2$. Find the marginal revenue $MR(Q)$ and the marginal cost $MC(Q)$ for the firm, and then find the firm’s profit-maximizing output. Does this quantity satisfy the first- and second-order conditions for the maximum?

-
- 1 More on the link between profit maximization and costs minimization is provided in the Appendix to Chapter 7.
 - 2 By looking at this figure demonstrating demand and total revenue, you may think again of Case 4.1 on the elasticity of demand of the iPhone.

CHAPTER 7

Costs and supply

Learning outcomes

By the end of this chapter, you should be able to:

- 1 describe a production function
- 2 recognize technology and a technique of production
- 3 understand how the choice of technique depends on input prices
- 4 analyse total, average and marginal cost, in the long run and short run
- 5 define returns to scale and their relation to average cost curves
- 6 recognize fixed and variable factors in the short run
- 7 describe the law of diminishing returns
8. understand how a firm chooses output, in the long run and short run

Chapter 6 introduced the theory of supply. Firms choose the output at which marginal cost equals marginal revenue. This maximizes profits (or minimizes losses). If profits are positive, the firm produces this output. If profits are negative, it checks whether losses are reduced by shutting down. This chapter develops the theory of supply in more detail. In particular, we want to understand how the firm can produce the output that maximizes profits (or minimizes losses). To do that, we need to better understand what the production activity of a firm is.

We distinguish between the *short-run* and the *long-run* output decisions of firms. No firm stays in business if it expects to make losses for ever. We show how and why cost curves differ in the short run, when the firm cannot fully react to changes in conditions, and the long run in which the firm can fully adjust to changes in demand or cost conditions.

Figure 7.1 summarizes the material of this chapter. The new material is all on the cost side. Because there are so many cost curves, you may find it useful to

check back to Figure 6.1. We start by introducing the *production function*, which describes the firm's technology.

7.1 Inputs and output: the production function

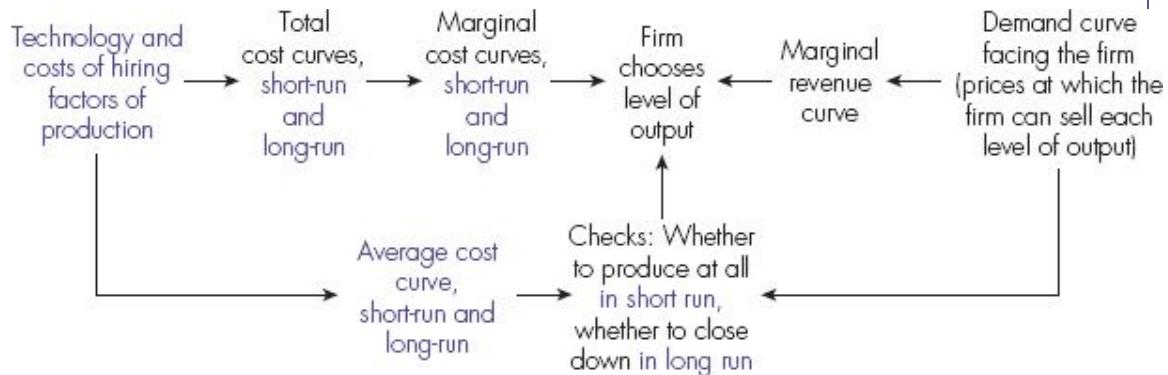
Suppose you want to start a farm to produce tomatoes. You need land, tomato seeds, water for irrigation, workers to work on the farm, a tractor and possibly some other machinery. All those are **inputs**, or **factors of production**, that are going to be used to produce tomatoes (output). Inputs include labour, machinery, buildings, raw materials and energy. The relationship between the quantity of inputs a firm uses and the output it produces is called a *production function*. We restrict the analysis to the case where only two inputs are used to produce a good or service. Those two inputs will be called *capital* (for example, machinery, buildings, and so on) and *labour* (for example, number of workers or number of worked hours).

An **input** (or **factor of production**) is a good or service used to produce output.

A production function is a way to summarize the technology available to the firm for its production activity in a given period of time. Suppose a firm uses inputs to make mobile phones. This is an engineering and management problem. Making mobile phones is largely a matter of technology and on-the-job experience. The **production function** summarizes **technically efficient** ways to combine inputs to produce output. Since profit-maximizing firms are not interested in wasteful production methods, we restrict our attention to those that are technically efficient.

Suppose our firm can use two different methods to produce the same number of mobile phones. To make 1 mobile phone, method A needs 2 workers and 1 machine, but method B needs 2 workers and 2 machines. Method B is less efficient than method A; it uses more machines but the same labour to make the same output. Method B is not in the production function.

A production technique is **technically efficient** if there is no other way to make a given output using less of one input and no more of the other inputs. The **production function** is the set of all technically efficient techniques.



This diagram extends the analysis of Chapter 6 in two ways. First, short-and long-run cost curves and output decisions are carefully distinguished. Second, we go behind the total cost curve to show how the firm chooses the lowest-cost way of producing each level of output, given the technology available to it and the costs of hiring factors of production.

Figure 7.1 The complete theory of supply

Short run vs long run

A firm's decisions about production activity depend on the time horizon. We distinguish between short run and long run. The *short run* is the period in which the firm cannot fully adjust to a change in conditions. In particular, we define the short run as a period of time in which the quantity of at least one input of production is **fixed**. A fixed input is a factor of production that cannot be increased or decreased in a given period of time. On the other hand, the *long run* is a period of time sufficiently long such that all inputs can be **varied**.

A **fixed factor of production** is an input that cannot be varied. A **variable factor** can be varied, even in the short run.

Consider a firm manufacturing cars. It has a production plant, which represents its capital input, and some workers, which represent its labour input. Suppose demand for the cars produced by the firm increases. This represents a change in the conditions faced by the firm. In the short run, the firm can increase production by increasing the number of workers in the existing production plant. In the long run, the firm has the possibility of building another production plant. However, such a change in its capital input requires time. The long run will be the period of time necessary for the firm to create a new production plant.

How long are the short and the long run? The answer depends on the specific industry. It might take ten years to build a new power station but only a few months to open new restaurant premises if an existing building can be bought, converted and decorated.

What really matters is that production decisions of firms can be different depending upon whether we take a short-run or long-run perspective.

7.2

Production in the short run: diminishing marginal returns

In the short run, the quantity of at least one input is fixed and cannot be varied. Capital will be the *fixed input* while labour will be the *variable input*.

In Table 7.1 we report the short-run production function for a hypothetical firm. In particular, the first two columns tell us how output produced rises as variable labour input is added to a fixed quantity of capital. When no workers are employed, given the existing fixed level of capital, output cannot be produced. When only 1 worker is employed, 2 units of output are produced. When 2 workers are employed, 6 units of output are produced, and so on.

The **marginal product** of a variable factor is the extra output from an extra unit of that input, holding constant all other inputs.

In the third column of Table 7.1 we have the **marginal product** of labour. This tells us how much an extra worker employed in the production process adds to the production activity, holding constant the level of capital. The first worker raises output by 2 units. We say that the first worker has a marginal product of 2 units. The second worker has a marginal product of 4 units. This is because, by adding that worker, the output produced rises from 2 to 6 units. The third worker has a marginal product of 8 units, since 2 workers produce 6 units but 3 workers produce 14 units.

Table 7.1 Total and marginal product of labour

Labour input (workers)	Output (total product per week)	Marginal product of labour
0	0	-
1	2	2
2	6	4
3	14	8

4	24	10
5	32	8
6	37	5
7	40	3
8	40	0
9	38	-2

The short-run production function outlined in Table 7.1, together with the marginal product of labour, is shown in Figure 7.2.

The top panel of Figure 7.2 shows total output as a function of labour input, keeping constant the level of capital. This is the total output curve.

Looking at that graph, we can see some important features. First, when few workers are employed, adding extra workers will increase the output produced more than proportionally. By increasing the number of workers, they can specialize more in what they do and the efficiency of the workers increases. This is the part of the graph between 0 and point A.

Second, after point A, adding extra workers will continue to increase the output produced but less than proportionally. When 4 workers are employed, total output is 24. Adding an extra worker will increase output to 32, but this extra worker is adding less to the production than the previous worker (8 compared to 10). The efficiency of the workers starts to decrease. This is because, in the short run, the level of capital (the other input of production) is fixed and cannot be changed.

Suppose the factory has 4 machines and there are 4 workers each specializing in fully running one of the factory's machines. If we add a fifth worker, her marginal product is lower. With only four machines, the fifth worker gets to use one only when another worker is having a rest. There is even less useful machine work for the sixth worker to do. In fact, beyond four workers, the marginal product of each additional worker decreases steadily as the number of workers is increased. We say that there are **diminishing marginal returns** to labour.

Holding all factors constant except one, the **law of diminishing marginal returns** says that, beyond some level of the variable input, further increases in the variable input lead to a steadily decreasing marginal product of that input.

The results behind Figure 7.2 are very general. Indeed, the idea of diminishing marginal returns from a variable input is a general law about

short-run technology. Adding ever more workers to a fixed quantity of machinery becomes less and less useful. The eighth worker's main role in production is to get coffee for the others. This does not contribute to output and we are at point B in Figure 7.2. Adding a ninth worker may also be harmful to production. In practice, the place is so crowded that the ninth worker is in the way of the others, disrupting them and so reducing the total output produced.

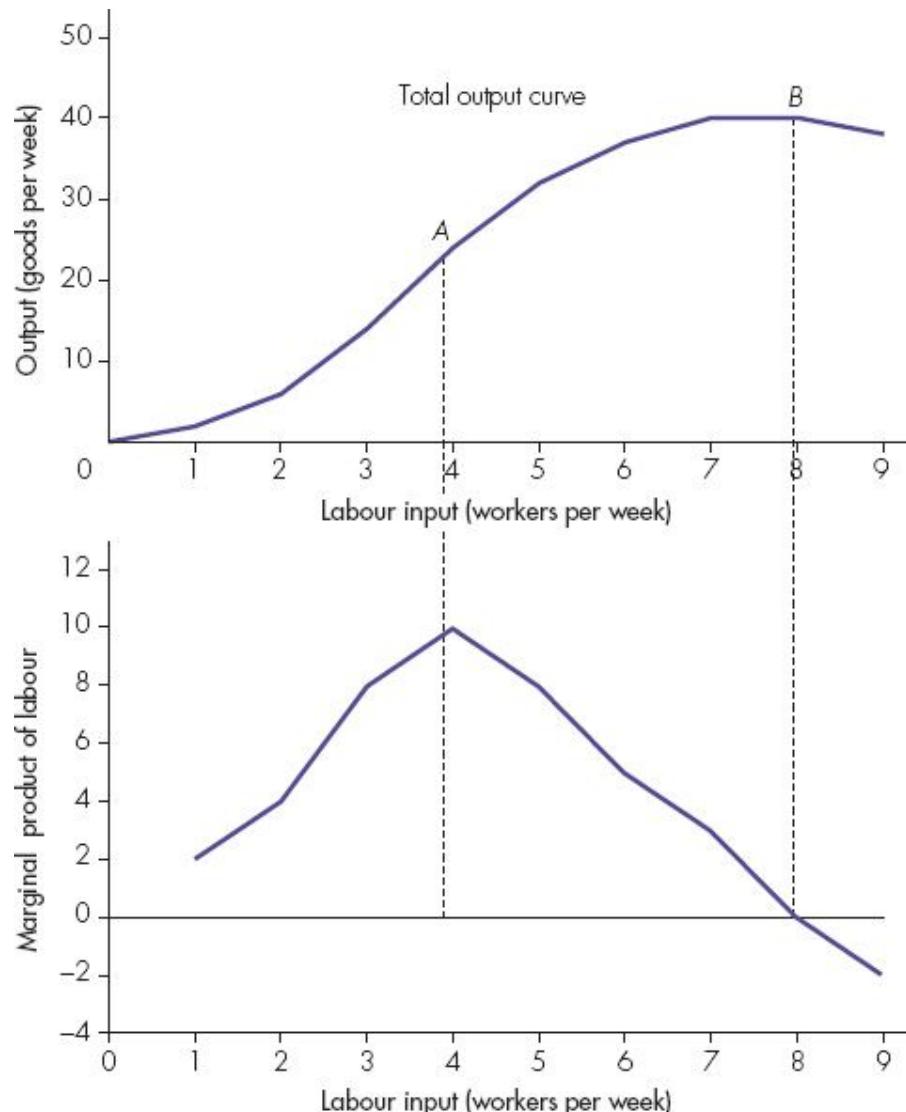


Figure 7.2 The productivity of labour and diminishing marginal returns

The diminishing marginal returns to labour are described by the bottom graph in Figure 7.2m. The relationship between the short-run total output curve and the marginal product of labour is the following: between 0 and point A on the total output curve, there are no diminishing marginal returns of labour. The

marginal product of labour is increasing with the number of workers. After point A in the total output curve, diminishing marginal returns take place and the marginal product of labour is decreasing with the number of workers. It must be stressed that the law of diminishing marginal returns of a variable input is a short-run phenomenon.¹

Notice that we have defined the marginal product of labour as the extra output obtained by employing an extra unit of labour, keeping constant the level of capital. So far we have considered a unit of labour as a single worker. However, we could have done the same analysis using an hour of work as a unit for the labour input. In that case, the marginal product of labour will be the extra output produced by adding an extra hour of labour.

If capital was the variable factor and labour the fixed factor, the result would be the same. Adding more and more machines to a given labour force might initially lead to large increases in output but would quickly encounter diminishing returns as machines become under-utilized. Thus the bottom graph in Figure 7.2, showing the marginal product of labour when labour is the variable factor, might also describe the behaviour of the marginal product of capital when capital is the variable factor.

Marginal product is *not* the everyday meaning of 'productivity', which refers to the average product. The average product of labour, what is most commonly meant by 'productivity', is total output divided by total labour input. The higher the productivity of labour, the higher is the output per worker produced.

If the marginal product of labour lies above the average product, adding another worker will raise the average product and 'productivity'. When diminishing returns set in, the marginal product will quickly fall below the average product and the latter will fall if further workers are added. If you do not see why this must be true, try calculating output per unit of labour input as an extra column in Table 7.1.

As usual, we must distinguish between movements along a curve and shifts in a curve. The marginal product curve is drawn for given levels of the other factors. For a higher given level of the fixed factors, the marginal product curve would be higher. With more machinery to work with, an extra worker will generally be able to produce more extra output than previously. The numbers in Table 7.1 and the height of the marginal product curve in Figure 7.2 depend on the number of fixed factors with which the firm began.

CASE 7.1

THE UK PRODUCTIVITY PUZZLE

In the text we said that labour productivity is measured as the average productivity of labour. This is just total output produced divided by the number of workers or by the number of hours worked, depending on how we measure labour input.

Between 2007 and 2009 UK real gross domestic product (GDP) fell by 5.8 per cent as a result of the credit crunch. Hours worked fell by 1.9 per cent, meaning workers were working less. Productivity fell by 3.9 per cent. This is intuitive because total output is falling more than the fall in hours worked and, therefore, workers must be less productive for that to happen.

However, between 2011 and the end of 2012 (latest available data), UK real GDP has grown by 1.3 per cent according to the data, but worker hours have increased by 2.3 per cent. UK workers are working more but they are less productive. Productivity decreased by 1 per cent. This situation is known as the 'UK productivity puzzle'. It is a puzzle because it is difficult to explain why workers work more but they have been less productive for almost two years. Did they lose their skills over the course of this long period? This seems unlikely.

A recent research paper shows that maybe this productivity puzzle is not a puzzle at all. We just measure productivity using the wrong data. In particular, we are neglecting in the calculation of output produced the role of intangible investments. They are intangible because they are not investments in physical goods like factories, machinery, and so on. Research and development (R&D) projects are typical intangible investments.

Here is how it works. Suppose a firm is reducing production but is investing resources in developing a new drug. This is an R&D project. This firm needs highly skilled and productive workers to work on the project. Therefore workers' productivity is high. The firm's measured output falls, since the output of the R&D project might not manifest itself for a few years. Thus measured productivity falls even if in reality it has increased.

Between 2011 and 2012 intangible investments increased substantially in the UK. Taking into account the growth of those investments in the calculation of real GDP, the research finds that labour productivity has instead increased.

Source: P. Goodridge, J. Haskel and G. Wallis, 'Can intangible investment explain the UK productivity puzzle?', Discussion Paper 2013/02, Imperial College Business School.

MATHS 7.1

THE SHORT-RUN PRODUCTION FUNCTION: THE AVERAGE AND MARGINAL PRODUCT OF LABOUR

The short-run production function can be written as:

$$Q = f(K_0, L) \quad (1)$$

where Q is the total output produced, K_0 is the fixed level of capital in the short run and L is the number of workers employed. Equation (1) simply says that the two inputs of production are combined through the function f in order to produce the output Q . The function f describes mathematically the technology available to the firm in the short run.

The average product of labour is defined as the average output produced by each worker. Mathematically, the average product of labour is defined as follows:

$$AP_L \equiv \frac{Q}{L} = \frac{f(K_0, L)}{L} \quad (2)$$

By now you should be familiar with the idea that marginal functions are defined using derivatives. Therefore the marginal product of labour is given by:

$$MP_L \equiv \frac{dQ}{dL} = \frac{df(K_0, L)}{dL} \quad (3)$$

Consider the following example for a short-run production function:

$$Q = K_0(L)^{0.5} \quad (4)$$

Equation (4) is a possible specification for the function f defined above. Suppose that $K_0 = 10$ and $L = 4$. The production function in (4) implies that the output produced by those 4 workers using 10 units of capital is 20.

The average product of labour implied by (4) is given by:

$$AP_L = \frac{K_0 L^{0.5}}{L} = \frac{K_0}{L^{0.5}} \quad (5)$$

If $K_0 = 10$ and $L = 4$, then the average product of labour according to (5) is 5. Each worker produces 5 units of output. The same result is obtained

if you divide the total output produced when $K_0 = 10$ and $L = 4$ by the number of workers (20/4 = 5).

Now suppose that $K_0 = 20$ instead. In this case, and with $L = 4$, the average product of labour is now 10. The workers are now more productive and each worker produces 10 units of output.

The marginal product of labour implied by (4) is given by:

$$MP_L = \frac{d(K_0 L^{0.5})}{dL} = 0.5 \frac{K_0}{L^{0.5}} \quad (6)$$

As you can see from equation (6), as labour increases, given the amount of capital K_0 , the marginal product of labour decreases, as suggested by the idea of diminishing returns to labour.

Equation (4) is an example of a short-run production function displaying decreasing marginal returns. We can think of other possible short-run production functions that do not have such a property. An example is given by the following linear short-run production function:

$$Q = K_0 L$$

In this case, the marginal product of labour is constant for any level of labour. Moreover, the marginal product of labour is always equal to the average product of labour.

7.3 Short-run costs

The short-run production function of a firm tells us the relationship between variable and fixed inputs and output produced. We can translate that information into a relationship between cost of production and output. The cost of production will depend on two elements. First, it depends on the price of inputs. A firm must pay a wage to the workers it employs and it must pay the price of machinery that it wants to buy. The higher the prices of inputs, the higher will be the cost of production. Second, the cost of production will depend on the productivity of the inputs. The higher is the productivity of inputs and the lower is the amount of inputs needed to produce a given level of output, the lower is thus the cost of producing that output.

The existence of fixed factors in the short run has two implications. First, in the short run the firm has some **fixed costs**. These fixed costs must be borne even if output is zero. If the firm cannot quickly add to or dispose of its

existing factory, it must still pay depreciation on the building and meet the interest cost of the money it originally borrowed to buy the factory.

Fixed costs do not vary with output.

Second, because in the short run the firm cannot make all the adjustments it would like, its short-run costs must exceed its long-run costs. We now study these short-run costs in more detail. Long-run costs will be discussed later in the chapter.

Short-run fixed and variable costs

Table 7.2 presents data on short-run costs to produce different levels of output. The second column shows the fixed costs, which are independent of the output level. These will be costs associated with the level of capital. The third column shows the variable costs. Output and all costs are measured per week.

Variable costs are the costs of hiring variable inputs, in our case labour. Firms may have long-term contracts with workers, which reduce the speed at which these inputs can be adjusted. Yet most firms retain some flexibility through overtime, short-term contracts and hiring casual or part-time workers.

Variable costs change as output changes.

The fourth column of Table 7.2 shows short-run total costs:

Short-run total cost (STC) = short-run fixed cost (SFC) + short-run variable cost (SVC)

The short-run total, fixed and variable cost curves are shown in Figure 7.3 using the data of Table 7.2. Notice that the shape of the short-run total cost curve is almost the mirror image of the shape of the short-run production function in Figure 7.2. This is not by accident but is a general result that is produced by the relationship between costs of production and the productivity of inputs.

Table 7.2 Short-run costs of production

(1) Output	(2) Short-run fixed cost (SFC)	(3) Short-run variable cost (SVC)	(4) Short-run total cost (STC)	(5) Short-run marginal cost (SMC)
0	30	0	30	-

1	30	22	52	22
2	30	38	68	16
3	30	48	78	10
4	30	61	91	13
5	30	79	109	18
6	30	102	132	23
7	30	131	161	29
8	30	166	196	35
9	30	207	237	41
10	30	255	285	48

Short-run marginal costs

The final column of Table 7.2 shows the **short-run marginal cost (SMC)**. Since fixed costs do not rise with output, SMC is the rise both in short-run total costs and short-run variable costs as output is increased by 1 unit.

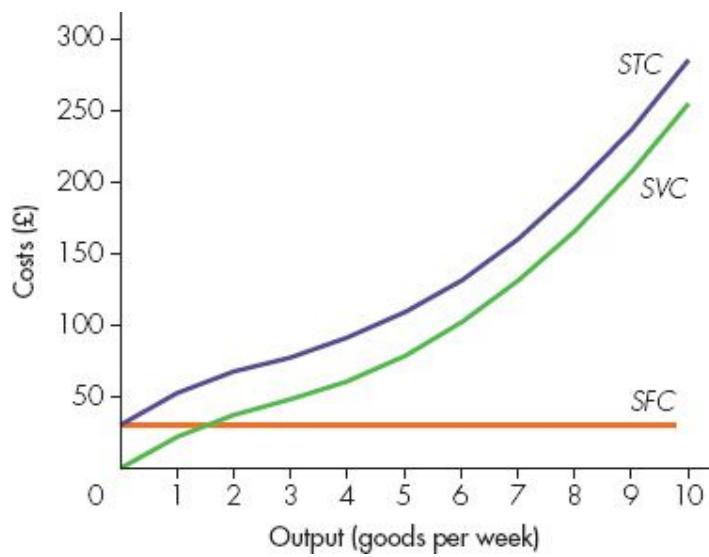
Short-run marginal cost (SMC) is the extra cost of making an extra unit of output in the short run while some inputs remain fixed.

Whatever the output, fixed costs are £30 per week. Marginal costs are always positive. Short-run total costs rise steadily as output rises. Extra output adds to total cost, and adds more the higher the marginal cost. In the last column of Table 7.2, as output increases, marginal costs first fall then rise again.

The short-run marginal cost curve is shown in Figure 7.4. As we can see, the shape in Figure 7.4 is almost the mirror image of the marginal product curve in Figure 7.2. Indeed, there is a close relationship between these two curves.

The short-run marginal cost is related to the variable input, in our case labour. Every worker costs the firm the same wage. While the marginal product of labour is increasing, each worker adds more to output than the previous workers. Hence the extra cost of making extra output is falling. SMC is falling as long as the marginal product of labour is rising.

Once diminishing returns to labour set in, the marginal product of labour falls and SMC starts to rise again. It takes successively more workers to make each extra unit of output. So, the shape of the short-run marginal cost curve is determined by the shape of the marginal product curve in Figure 7.2, which in turn depends on the technology facing the firm.



The short-run total cost (STC) is the sum of the short-run fixed cost (SFC) and the short-run variable cost (SVC). The short-run fixed cost does not depend on the level of output and therefore its curve is a horizontal line.

Figure 7.3 Short-run total, fixed and variable cost curves

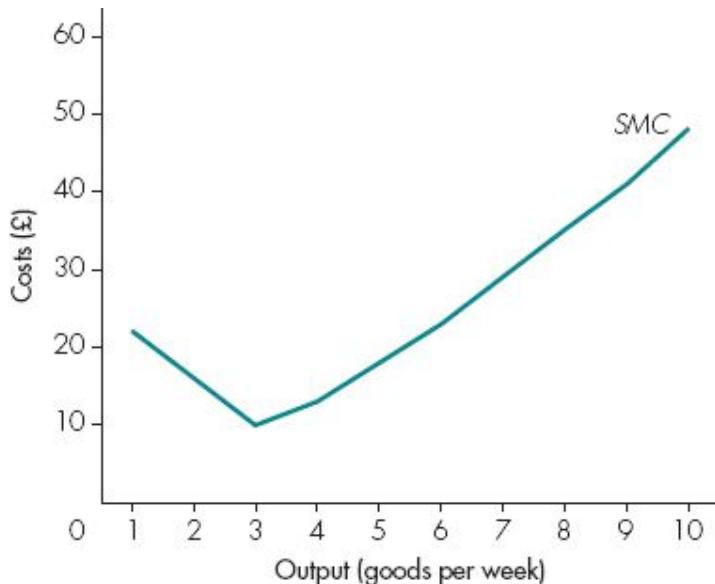


Figure 7.4 Short-run marginal cost curve

The short-run marginal cost (SMC) curve is first decreasing with output and then increases as output continues to increase. To stress that marginal cost is incurred by moving from one output level to another, we plot SMC at points halfway between the corresponding outputs.

Short-run average costs

Another important cost measure is represented by the average cost. The average cost is defined as the total cost divided by the quantity produced. Therefore, the average cost measures the total cost per unit of output.

The short-run average cost is given by:

Short-run average total cost (SATC) = short-run total cost (STC)/Quantity of output

In the short run, the total cost is given by the sum of variable and fixed costs. Therefore, together with the short-run total average cost, we can also define the short-run variable cost and the short-run fixed cost.

Short-run average variable cost (SAVC) = short-run variable cost (SVC)/Quantity of output

Short-run average fixed cost (SAFC) = short-run fixed cost (SFC)/Quantity of output

Using those two average cost measures, we can define the short-run average total cost in the following way, equivalent to equation (2):

Short-run average total cost (SATC) = short-run average fixed cost (SAFC) + short-run average variable cost (SAVC)

This follows from dividing each term in equation (1) by the output level.

Table 7.3 shows short-run *average* cost data corresponding to Table 7.2. Each number in Table 7.3 is obtained by dividing the corresponding number in Table 7.2 by the output level. The table also shows short-run marginal costs, taken from Table 7.2.

Figure 7.5 plots the three short-run average cost measures from Table 7.3.

In Figure 7.5 SAFC falls steadily because total fixed cost ('overheads') is spread over ever-larger output levels, thus reducing average fixed cost. The SATC and SAVC curves are such that, at each output level, SATC = SAVC + SAFC, as in equation (3).

Short-run average variable cost (SAVC) equals SVC divided by output, and short-run average total cost (SATC) equals STC divided by output.

Short-run average fixed cost (SAFC) equals short-run fixed cost (SFC) divided by output.

Table 7.3 Short-run average costs of production

Output	Short-run average fixed cost (SAFC)	Short-run average variable cost (SAVC)	(Short-run average total cost (SATC))	Short-run marginal cost (SMC)
1	30.00	22.00	52.00	22
2	15.00	19.00	34.00	16
3	10.00	16.00	26.00	10
4	7.50	15.25	22.75	13
5	6.00	15.80	21.80	18
6	5.00	17.00	22.00	23
7	4.29	18.71	23.00	29
8	3.75	20.75	24.50	35
9	3.33	23.00	26.33	41
10	3.00	25.50	28.50	48

On the relationship between short-run marginal and average costs

In Figure 7.6 we plot the short-run marginal and average costs. Two facts stand out from this figure:

1. $SATC$ is falling when SMC is less than $SATC$, while it is rising when SMC is greater than $SATC$. The same applies for the relationship between SMC and $SAVC$.
2. $SATC$ is at a minimum at the output at which the SMC curve and the $SATC$ curve cross (point A in Figure 7.6). The $SAVC$ is at its minimum at the output at which the SMC curve and the $SAVC$ curve cross (point B in Figure 7.6).

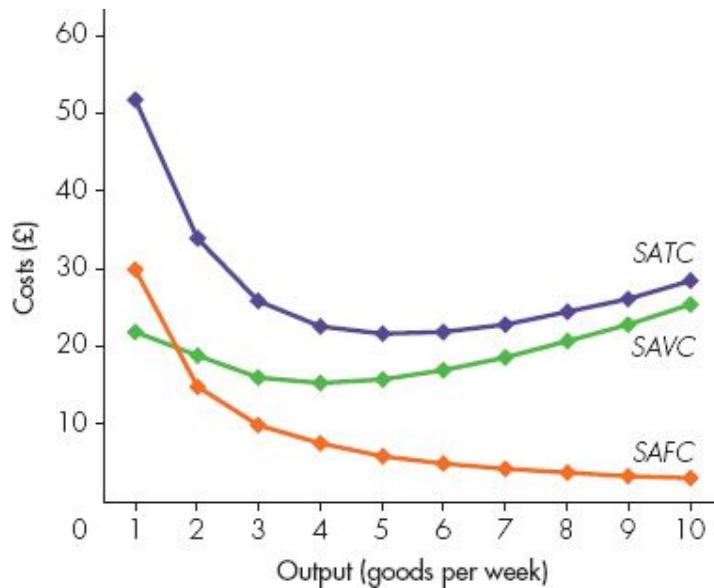


Figure 7.5 Short-run average cost curves

Neither fact is an accident. The relationship between average and marginal is a matter of arithmetic, as relevant for football as for production costs. A footballer with 3 goals in 3 games averages 1 goal per game. Two goals in the next game, implying 5 goals from 4 games, raise the average to 1.25 goals per game. In the fourth game, the marginal goals were 2, raising total goals from 3 to 5. Because the marginal score exceeds the average score in previous games, the extra game must drag up the average.

The same holds for production costs. When the marginal cost of the next unit exceeds the average cost of the existing units, making the next unit must raise average cost. If the marginal cost of the next unit lies below the average cost of existing units, an extra unit of production drags down average costs. When marginal and average costs are equal, adding a unit leaves average cost unchanged. This explains fact 1.

Fact 2 follows from fact 1. In Figure 7.6 the short-run total average and marginal cost curves cross at point *A*, which must be the minimum point for *SATC*. To the left of *A*, *SMC* is below *SATC* so the short-run average total cost is still falling. To the right of *A*, *SMC* is above *SATC* so the short-run average total cost is rising. The short-run average total cost is lowest at *A*. The short-run marginal cost curve crosses the *SATC* curve from below, at the point where *SATC* is at the minimum. As in the football example, this rests purely on arithmetic. The same reasoning can be applied to the relationship between *SMC* and *SAVC*.

Table 7.4 summarizes this important relationship. It is true for the relationship between marginal and average costs both in the short run and in the long run.

Table 7.4 Marginal and average cost

	MC < AC	MC = AC	MC > AC
AC is:	Falling	Minimum	Rising

The shape of the SMC curve in Figure 7.6 follows from the behaviour of marginal labour productivity. The SMC curve passes through the lowest point, A , on the short-run average total cost curve. To the left of this point, SMC lies below $SATC$ and is dragging it down as output expands. To the right of A , the converse holds. That explains the shape of the $SATC$ curve in Figure 7.6 and in Figure 7.5 as well.

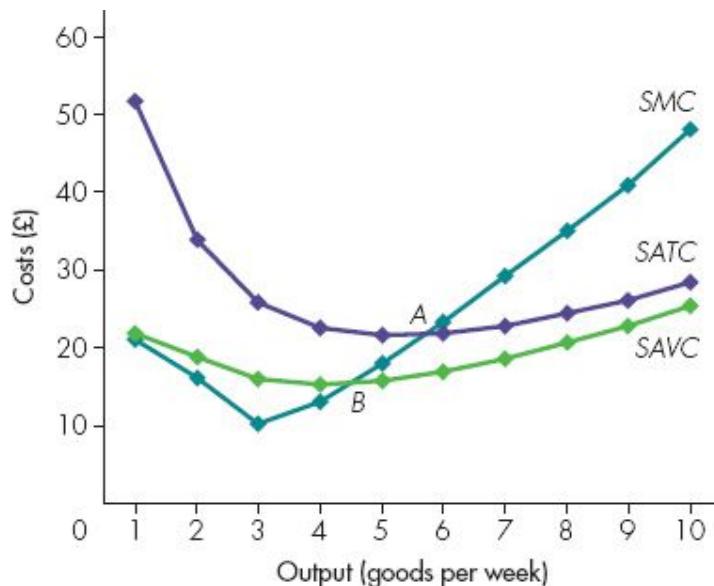


Figure 7.6 Short-run marginal and average cost curves

Variable cost is total cost minus fixed cost. Fixed cost does not change with output. Hence marginal cost also shows how much total *variable* cost is changing. The SMC curve goes through the lowest point, B , on $SAVC$. To the left of B , SMC is below $SAVC$ and $SAVC$ is falling. To the right of B , $SAVC$ is rising. Finally, since average total cost exceeds average variable cost by average fixed cost, $SAVC$ lies below $SATC$. Point B must lie to the left of point A . That explains the shape of $SAVC$ and its relation to $SATC$ in Figure 7.6 and in Figure 7.5 as well.

MATHS 7.2

SHORT-RUN COST FUNCTIONS

A short-run cost function represents the minimum cost to produce a given level of output when some factors of production cannot be adjusted.² An example of a short-run total cost function is given by:

$$STC = F + cQ + dQ^2 \quad (1)$$

The term F is a constant and denotes the short-run fixed cost that does not depend on output Q .

The term $cQ + dQ^2$ denotes the short-run variable cost that varies with output Q .

Therefore, from (1) we have: $SFC = F$ and $SVC = cQ + dQ^2$

The short-run marginal cost is measured by the change in STC as Q changes:

$$SMC \equiv \frac{dSTC}{dQ} = c + 2dQ \quad (2)$$

The short-run average fixed cost is given by:

$$SAFC \equiv \frac{SFC}{Q} = \frac{F}{Q} \quad (3)$$

The short-run average fixed cost decreases steadily as Q increases.

The short-run average variable cost is given by:

$$SAVC \equiv \frac{SVC}{Q} = c + dQ \quad (4)$$

The short-run average total cost is given by:

$$SATC \equiv \frac{STC}{Q} = \frac{F}{Q} + c + dQ \quad (5)$$

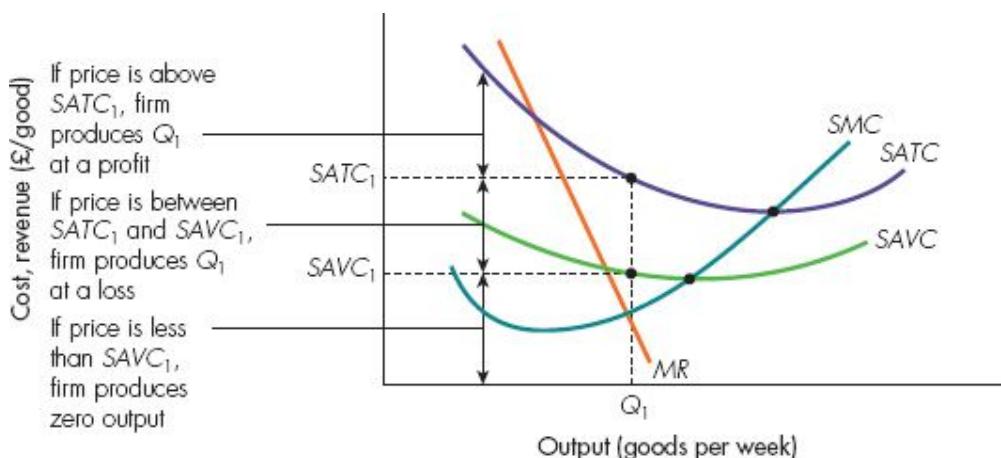
7.4

A firm's output decision in the short run

The firm's **short-run output decision** is to supply Q_1 , the output at which $MR = SMC$, if the price covers short-run average variable cost $SAVC_1$ at that output. If not, the firm supplies zero.

Figure 7.7 illustrates the firm's **short-run output decision**. Short-run marginal cost is set equal to marginal revenue to determine the output Q_1 that maximizes profits or minimizes losses.

Next, the firm decides whether or not to produce in the short run. Profit is positive at the output Q_1 if the price p at which this output is sold covers average total cost. It is the short-run measure $SATC_1$ at output Q_1 that is relevant. If p exceeds $SATC_1$, the firm makes profits in the short run and produces Q_1 .



The firm sets output at Q_1 , where short-run marginal costs equal marginal revenue. Then it checks whether it should produce at all. If price is above $SATC_1$, the level of short-run average total cost at output Q_1 , the firm is making a profit and should certainly produce Q_1 . If price is between $SATC_1$ and $SAVC_1$, the firm partly covers its fixed costs, even though it is losing money. It should still produce output Q_1 . Only if the price is below $SAVC_1$ should the firm produce zero. At those prices, the firm is not even covering its variable costs.

Figure 7.7 The firm's short-run output decision

Suppose p is less than $SATC_1$. The firm is losing money because p does not cover costs. In the long run the firm closes down if it keeps losing money. In the short run, even at zero output the firm must pay its fixed costs. The firm needs to know whether losses are bigger if it produces at Q_1 or produces zero.

If revenue exceeds *variable cost*, the firm is earning something towards its overheads. It produces Q_1 if revenue exceeds variable cost even though Q_1

may involve losses. The firm produces Q_1 if p exceeds $SAVC_1$. If not, it produces zero. Table 7.5 summarizes the short-run output decisions of a firm.

Table 7.5 The firm's output decision in the short run

	Marginal condition	Check whether to produce
Short run	Choose the output at which $MR = SMC$	Produce this output if $p > SAVC$ Otherwise, produce zero

ACTIVITY 7.1

MARGINAL CONDITIONS AND SUNK COSTS

The theory of supply obeys two principles of good decision making in life. The first is the *marginal principle*. Once the best position is reached, no feasible change can improve things. To climb a hill, take small steps in an upward direction. If you cannot move upward, you are at the top.³

There is also the big picture. Having equated marginal cost and marginal revenue, a firm checks whether it is not better to close down completely. Similarly, the marginal principle guides you to a local peak but, looking around, you may see a higher hill a mile away, but you have to go down a bit before you can scale it.

The second principle is that *sunk costs are sunk*. Costs already incurred should not affect new decisions. In choosing short-run output, the firm ignores fixed costs that are paid anyway. It is no use crying over spilt milk. Having read seven chapters of this book, should you read on? It depends on the costs and benefits you get from the rest of the book, not on the time already spent.

Questions

- a. A firm operates for two periods and then dies. In the first period, it can choose to buy a very special piece of equipment that will be no use to any other firm and will have no resale value. It will, however, help the firm to make output in each of the two periods. When the second period arrives, should the cost of the machine be included in the marginal costs of the firm? In the first period, should it be included in the marginal cost of producing output? What is the smart way for the firm to think about this problem?
- b. Playing poker, you bet most of your chips on a single hand before getting a sinking feeling that you are going to lose the hand. Should

you bet on? Why or why not?

To check your answers to these questions, go to page 678.

7.5 Production in the long run

In the long run all factors of production are variable. Table 7.6 shows some technically efficient methods in the production function. The first two rows show two ways to make 100 mobile phones: 4 machines and 4 workers, or 2 machines and 6 workers. Beginning from the latter, the third row shows the effect of adding an extra worker. Output rises by 6 mobile phones. The last row shows that doubling both inputs in the second row also doubles the output, though this need not be so: overcrowding a small factory can slow people down.

Table 7.6 A production function

Output	Capital input	Labour input
100	4	4
100	2	6
106	2	7
200	4	12

In the long run, the problem facing the firm is choosing the right mix of inputs to produce the quantity that maximizes the firm's profit. In practice, among all the possible available efficient techniques the firm must choose the cheapest one.⁴

Technical progress is a new technique allowing a given output to be made with fewer inputs than before.

A method that was previously technically efficient may become inefficient after a technical advance allows for a better production technique. **Technical progress** alters the production function. For now, we assume a given technology and a given production function. Chapter 28 discusses growth and technical progress.

Costs and the choice of technique

Consider the lowest-cost way to make 100 mobile phones.⁵ Assume there are two technically efficient techniques: the first two rows of Table 7.6, reproduced as the second and third columns of Table 7.7 and labelled techniques A and B. It costs £320 to rent a machine and £300 to hire a worker.

To make 100 mobile phones, Table 7.7 shows that the total cost is £2480 with technique A and £2440 with technique B. The firm chooses B. One hundred mobile phones at a total cost of £2440 is one point on the total cost curve for mobile phones. It is the *economically efficient* (lowest-cost) production method at the rental and wage rates in Table 7.7.

Table 7.7 Choosing the lowest-cost production technique

Technique	Capital input	Labour input	Rental per machine (£)	Wage per worker (£)	Capital cost (£)	Labour cost (£)	Total cost (£)
A	4	4	320	300	1280	1200	2480
B	2	6	320	300	640	1800	2440

To get the total cost curve, we repeat the calculation for each output. The production function tells us the inputs needed by each technique. Using input prices, we calculate the cost using each technique and choose the lowest-cost production method. Joining up these points, we get the total cost curve, which may switch from one production technique to another at different outputs. From the total cost curve we calculate the marginal cost curve; that is, the rise in total cost at each output when output is increased by one more unit.

Factor intensity

A technique using a lot of capital and little labour is ‘capital intensive’. One using a lot of labour but relatively little capital is ‘labour intensive’. In Table 7.7, technique A is more capital intensive and less labour intensive than technique B. The ratio of capital input to labour input is 1 in technique A but only $\frac{1}{3}$ in technique B.

Factor prices and the choice of technique

At the factor prices (prices per unit input) in Table 7.7, the more labour-intensive technique is cheaper. Suppose the wage rises from £300 to £340: labour is dearer but the rental on capital is unchanged. The *relative price* of labour has risen.

We ask two questions. First, what happens to the total cost of making 100 mobile phones? Second, is there any change in the preferred technique? Table

7.8 recalculates production costs at the new factor prices. Because both techniques use some labour, the total cost of making 100 mobile phones by each technique rises. Repeating this argument at all outputs, the total cost curve must shift *upwards* when the wage rate (or the price of any other input) rises.

In this example, the rise in the relative price of labour leads the firm to switch techniques: it switches to the more capital-intensive technique, A.

Table 7.8 Effect Of a higher wage rate

Technique	Capital input	Labour input	Rental per machine (£)	Wage per worker (£)	Capital cost (£)	Labour cost (£)	Total cost (£)
A	4	4	320	340	1280	1360	2640
B	2	6	320	340	640	2040	2680

7.6 Long-run total, marginal and average costs

Faced with an upward shift in its demand and marginal revenue curves, a firm will expand output, as we explained in Chapter 6. However, adjustment takes time. Initially, the firm can get its existing workforce to do overtime. In the long run, the firm can vary its factory size, switch techniques of production, hire new workers and negotiate new contracts with suppliers of raw materials.

Long-run total cost (LTC) is the minimum cost of producing each output level when the firm can adjust all inputs.

Long-run marginal cost (LMC) is the rise in long-run total cost if output rises permanently by one unit.

Long-run average cost (LAC) is the total cost LTC divided by the level of output Q.

The firm may be able to alter the shift length at once. Hiring or firing workers takes longer and it might be years before a new factory is designed, built and operational. In this section we deal with long-run cost curves, when the firm can make all the adjustments it desires.

Table 7.9 shows the **long-run total cost (LTC)** and **long-run marginal cost (LMC)** of making each output. Since there is always an option to close down

entirely, the *LTC* of producing zero output is zero. *LTC* describes the eventual cost after all adjustments have been made.

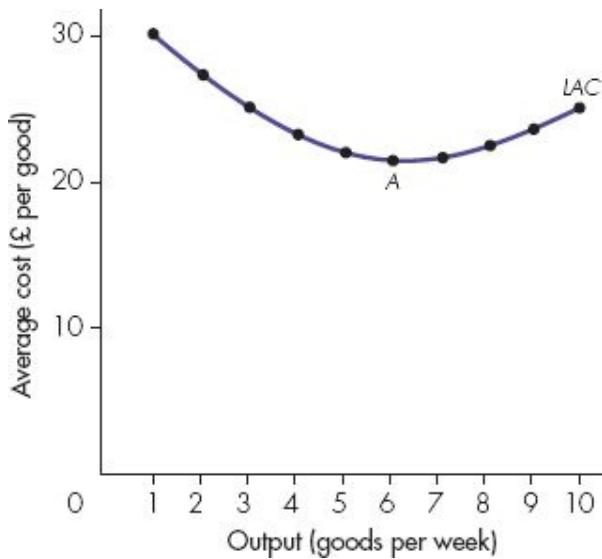
Table 7.9 also shows long-run marginal cost *LMC*. *LTC* must rise with output: higher output always costs more to produce. *LMC* shows how much total cost is involved in making the last unit of output.

Can large firms produce goods at a lower unit cost than small firms? Might it be a disadvantage to be large? To answer these questions, we need to think about average cost per unit of output.

Table 7.9 shows **long-run average cost (*LAC*)** (column 2 divided by column 1). These *LAC* data are plotted in Figure 7.8. Average cost starts out high, then falls, then rises again. This is a similar shape to the short-run average cost function previously considered. This common pattern of average costs is called the U-shaped average cost curve. To see why the U-shaped average cost curve is common in practice, we examine ‘returns to scale’.

Table 7.9 Long-run costs

(1) Output	(2) Total cost (£)	(3) Marginal cost (£)	(4) Average cost (£)
0	0	–	–
1	30	30	30
2	54	24	27
3	74	20	24.67
4	91	17	22.75
5	107	16	21.40
6	126	19	21.00
7	149	23	21.29
8	176	27	22.00
9	207	31	23.00
10	243	36	24.30

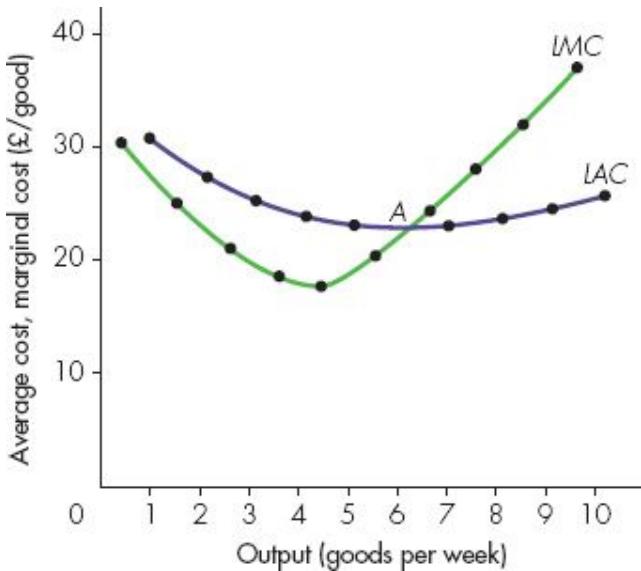


This long-run average cost curve LAC plots the data in the last column of Table 7.9. The LAC curve has the typical U-shape. The minimum average cost of production is at point A, with an output level of 6 and average cost of £21.

Figure 7.8 The LAC curve

The last two columns of Table 7.9 are plotted in Figure 7.9. At each output, *LAC* is total cost divided by output. To stress that marginal cost is incurred by moving from one output level to another, we plot *LMC* at points halfway between the corresponding outputs. The *LMC* of £30 for the first unit of output is plotted at the output halfway between 0 and 1.

The relationship between *LMC* and *LAC* is exactly the same as the one discussed for the short-run case. When *LMC* is below *LAC*, *LAC* decreases. When *LMC* is above *LAC*, *LAC* increases. Furthermore, *LMC* and *LAC* cross at the point where *LAC* reaches a minimum.



These cost data are plotted from Table 7.9. There are two special features of the relationship between the marginal cost curve (LMC) and the average cost curve (LAC). First, LAC is declining whenever LMC is below LAC, and rising whenever LMC is above LAC. Second, the LMC curve cuts the LAC curve at the minimum point of the LAC curve – in other words, at the point where output is produced at lowest unit cost.

Figure 7.9 Long-run average and marginal cost curves

7.7 Returns to scale

Scale refers to the output of the firm when all inputs can be varied. Therefore it is a long-run concept. Returns to scale refer to the relationship between the long-run average cost and output produced by a firm when all inputs of production are variable. When long-run average cost decreases as output increases, the firm faces *increasing returns to scale* (or economies of scale). When the long-run average cost remains constant as output increases, the firm faces *constant returns to scale*. If the long-average cost increases with output, then we have *decreasing returns to scale* (or diseconomies of scale). The three cases are shown in Figure 7.10.

In Figure 7.8 the U-shaped average cost curve had economies of scale up to point *A*, where average cost was lowest. At higher outputs there were diseconomies of scale. Why are there scale economies at low output levels but diseconomies of scale at high output levels?

We draw a cost curve for given input prices. Changes in average cost as we move along the *LAC* curve cannot be explained by changes in factor prices. (Changes in factor prices *shift* cost curves.) The relationship between average

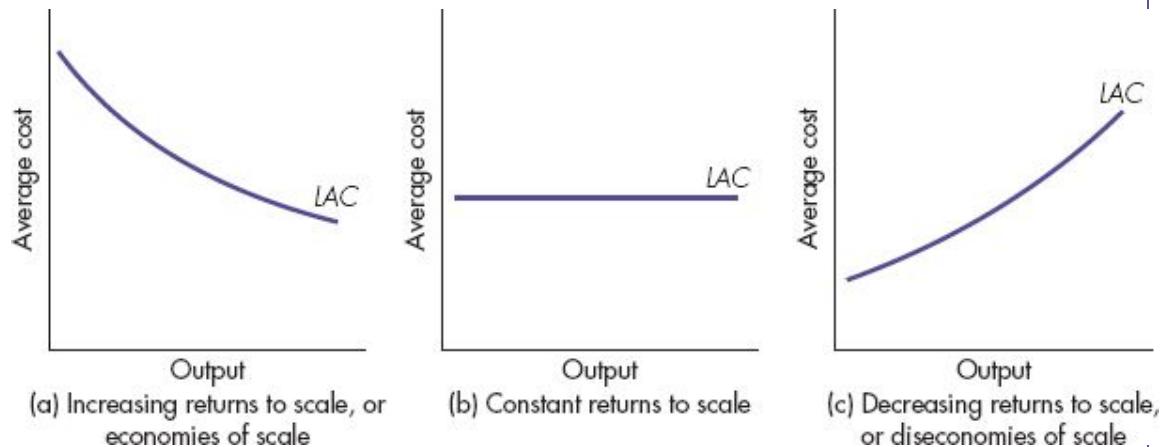
cost and the output LAC curve depends on the technical relation between physical quantities of inputs and output, summarized in the production function.

Economies of scale

There are three reasons for economies of scale. The first is *indivisibilities* in the production process; that is, a minimum quantity of inputs required by the firm to be in business at all whether or not output is produced. These are sometimes called *fixed costs*, because they do not vary with the output level. To be in business a firm requires a manager, a telephone, an accountant and a market research survey. The firm cannot have half a manager and half a telephone merely because it wishes to operate at low output levels.

Economies of scale (or **increasing returns to scale**) mean long-run average cost falls as output rises.

Beginning from small output levels, these costs do not initially increase with output. The manager can organize three workers as easily as two. As yet there is no need for a second telephone. There are economies of scale because these fixed costs can be spread over more units of output as output is increased, reducing average cost per unit of output. However, as the firm expands further, it has to hire more managers and telephones and these economies of scale die away. The average cost curve stops falling.



The three long-run average cost LAC curves show the relationship between returns to scale and the shape of the LAC curve. When LAC is declining, average cost of production falls as output increases and there are economies of scale. When LAC is increasing, average cost of production increases with higher output, and there are decreasing returns to scale. The intermediate case, where average cost is constant, has constant returns to scale.

Figure 7.10 Returns to scale and long-run average cost curve

The second reason for economies of scale is *specialization*. A sole trader must undertake all the different tasks of the business. As the firm expands and takes on more workers, each worker can concentrate on a single task and handle it more efficiently.

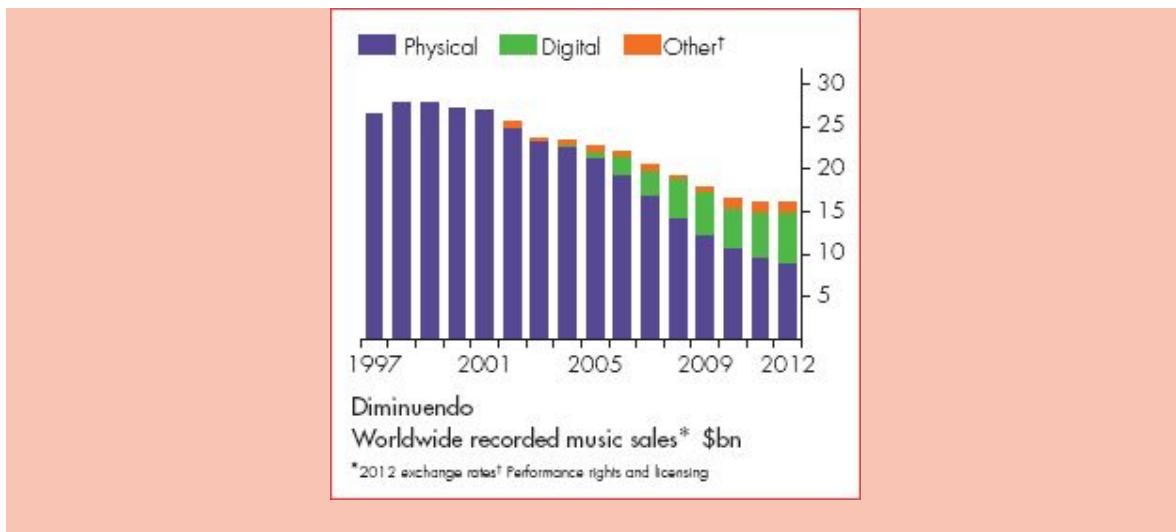
The third reason for economies of scale is closely related. Large scale is often needed to take advantage of better machinery. No matter how productive a robot assembly line is, it is pointless to install one to make five cars a week. Average costs would be enormous. However, at high output levels the machinery cost can be spread over a large number of units of output and this production technique may produce so many cars that average costs are low.

CASE 7.2

SCALE ECONOMIES AND THE INTERNET

Producing information products such as films, music and news programmes has a high fixed cost, but distributing these products digitally has almost a zero marginal cost and no capacity constraint. Scale economies are vast. Moreover, if marginal cost is close to zero, smart suppliers will price their products so that marginal revenue is also tiny.

The Internet had a big negative impact on the music industry, but it is now helping it resurface. People can now buy high quality music tracks on iTunes for £0.99 and lower quality tracks for £0.79. Digital sales rose by 9 per cent in 2012 and a third of the music industry's revenues now come through digital channels. Subscription services on popular channels like Spotify and Deezer, which pay royalties each time a song is played, increased by 44 per cent in 2012. The increasing use of Smartphones has made downloading and streaming music legally more attractive.



Source: 'Something to sing about', © The Economist Newspaper Limited, London, 2 Mar 2013.

Diseconomies of scale

Diseconomies of scale (or **decreasing returns to scale**) mean long-run average cost rises as output rises.

Beyond some output, the U-shaped average cost curve turns up again as diseconomies of scale begin. Management is harder as the firm gets larger: there are *managerial diseconomies of scale*. Large companies need many layers of management, themselves needing to be managed. The company becomes bureaucratic, co-ordination problems arise and average costs begin to rise.

Geography may also explain diseconomies of scale. If the first factory is located in the best site, to minimize the cost of transporting goods to the market, the site of a second factory must be less advantageous. To take a different example, in extracting coal from a mine, a firm will extract the easiest coal first. To increase output, deeper coal seams have to be worked and these will be more expensive.

As output increases, the shape of the average cost curve thus depends on two things: how long economies of scale persist and how quickly the diseconomies of scale set in. The balance of these two forces varies from industry to industry and from firm to firm.

THE LONG-RUN PRODUCTION FUNCTION AND THE RETURNS TO SCALE

We defined the returns to scale in terms of the relationship between the long-run average cost and the level of output. We can define the same concepts using the relationship between inputs and output implied by a long-run production function.

We say that a production function displays increasing, constant or decreasing returns to scale if the following definitions hold:

- a. *Increasing returns to scale (or economies of scale)*: when all the inputs of production are increased by the same factor and the output produced increases more than proportionally. Constant returns to scale mean long-run average costs are constant as output rises.
- b. *Constant returns to scale* : when all the inputs of production are increased by the same factor and the output produced increases by the same factor.
- c. *Decreasing returns to scale (or diseconomies of scale)*: when all the inputs of production are increased by the same factor and the output produced increases less than proportionally.

Definition (a) is equivalent to saying that the long-run average cost is decreasing with output. Suppose we double the amount of inputs used in production. The long-run total cost will double as well. Output produced increases by more than double according to definition (a). Since the long-run average cost is the long-run total cost divided by the output, it implies that the *LAC* must decrease after we double all the inputs. Similar reasoning can be used to explain why definitions in (b) and (c) are equivalent to the ones given in the text.

Mathematically, consider the following long-run production function, known as a Cobb – Douglas production function:

Constant returns to scale mean long-run average costs are constant as output rises.

$$Q = K^\alpha L^\beta \quad (1)$$

where $\alpha > 0$ and $\beta > 0$ are two constants.

To apply the definitions written above, we need to increase all the inputs of production by the same factor. Suppose we increase all inputs by 1 . 0. What happens to the output produced?

The long-run production function in (1) becomes:

$$Q_1 = (\lambda K)^\alpha (\lambda L)^\beta \quad (2)$$

where Q_1 denotes the output produced when all inputs are increased by the same factor 1.0. A bit of algebra to simplify equation (2) gives us:

$$Q_1 = \lambda^\alpha K^\alpha \lambda^\beta L^\beta = \lambda^{\alpha+\beta} K^\alpha L^\beta$$

Using the fact that $Q = K^\alpha L^\beta$, this can be rewritten as:

$$Q_1 = \lambda^{\alpha+\beta} Q$$

Equation (3) tells us that, when we increase inputs by the same factor 1.0, the output produced increases by a factor $\lambda^{\alpha+\beta}$.

Then we have the following:

1. If $\alpha + \beta > 1$, the Cobb – Douglas production function displays increasing returns to scale.
2. If $\alpha + \beta = 1$, the Cobb – Douglas production function displays constant returns to scale.
3. If $\alpha + \beta < 1$, the Cobb – Douglas production function displays decreasing returns to scale.

If a firm is using a production function that displays increasing returns to scale, then the long-run average cost of that firm will decrease with output.

Returns to scale in practice

To gather evidence on returns to scale, we can talk to design engineers to see how production costs vary with output. It is much harder to quantify managerial diseconomies. Most empirical research focuses only on direct production costs. Because it ignores managerial diseconomies of scale, it overestimates scale economies.

Many such studies of manufacturing industry confirm that scale economies continue over a wide range of output.⁶ The long-run average cost curve slopes down, albeit at an ever-decreasing rate. Economists have tried to measure the output at which all scale economies are first achieved: the point at which the average cost curve first becomes horizontal.

Minimum efficient scale (MES) is the lowest output at which the LAC curve reaches its minimum.

Table 7.10 contains some estimates of the **minimum efficient scale (MES)** for firms in different industries in the UK and the US. The second column gives an idea of how steeply average costs fall before minimum efficient scale is reached. It shows how much average costs are higher if output is one-third the output of minimum efficient scale. The third and fourth columns show the *MES* output relative to the output of the industry as a whole. This provides a benchmark for the importance of economies of scale to firms in each industry. Since firms in the UK and the US have access to essentially the same technical know-how, differences between the third and fourth columns primarily reflect differences in the size of the industry in the two countries rather than differences in the *MES* output level for an individual firm.

Table 7.10 Minimum efficient scale selected industries, UK and US

Industry	% increase in LAC at $\frac{1}{3}$ MES	MES as % of market in	
		UK	US
Cement	26	6	2
Steel	11	15	3
Glass bottles	11	9	2
Bearings	8	4	1
Fabrics	7	2	1
Refrigerators	6	83	14
Petroleum	5	12	2
Paints	4	10	1
Cigarettes	2	30	6
Shoes	2	1	1

Source: F. M. Scherer et al., *The economics of multiplant operation*, Harvard University Press, 1975, Tables 3.11 and 3.15.

Scale economies in manufacturing industries are substantial. At low outputs, average costs are much higher than at minimum efficient scale. We would expect similar effects in aircraft and motor car manufacture, which have huge fixed costs for research and development of new models and which can utilize highly automated assembly lines once output is large. Yet in a large country such as the US, minimum efficient scale for an individual firm occurs at an output that is small relative to the industry as a whole. Most firms are producing on a relatively flat part of their average cost curve, with few scale economies unexploited. In smaller countries such as the UK, the point of minimum efficient scale is larger relative to the industry as a whole.

However, Table 7.10 suggests that there are many industries, even in the manufacturing sector, where minimum efficient scale for a firm is small relative to the whole market and average costs are only a little higher if output is below minimum efficient scale. These firms will be producing in an output range where the *LAC* curve is almost horizontal. Finally, there are many firms, especially outside the manufacturing sector, whose cost conditions are well represented by a U-shaped average cost curve. With only limited opportunities for economies of scale, these firms run into rising average costs even at quite moderate levels of output. Many service-sector industries – hairdressers, doctors, decorators – have very modest scope for scale economies.

Globalization, technical change and scale economies

The estimates in Table 7.10 refer to data more than a decade old. Some things have changed since then. Technical progress in transport has reduced the cost of shipping goods over vast distances. Technical progress in information technology has made it much easier to manage companies with global activities. Computers at courier companies such as Federal Express and DHL can track packages across the world. In 2001 the global activities of FedEx were temporarily halted not by a pilots'

Globalization is the increasing integration of national markets that were previously much more segmented from one another.

Globalization is partly a matter of policy – countries are abandoning restrictions to keep out foreign businesses – but it is chiefly being driven by cost changes caused by technical progress. New technology and lower transport costs do not merely enhance market size, they also reduce managerial diseconomies of scale. It gets easier to run big companies. The output of minimum efficient scale is rising. Global companies like Microsoft, Shell, Nike and Nokia keep popping up successfully in more and more countries: scale economies let them undercut the domestic competition.

Globalization is the increasing integration of national markets that were previously much more segmented from one another.

The second sense in which the presentation of Table 7.10 is ‘traditional’ is that its final column presumes that the domestic market size is the relevant market size against which to assess minimum efficient scale. That would

make sense if firms produced only for the home market. Globalization is making this obsolete, too. Of course, the larger the potential market, the easier it is to justify large scale, and the more firms may seek mergers in order to achieve that scale quickly.

For example, when Barclays bank announced in 2007 its proposed £45 billion takeover of Dutch counterpart ABN Amro, it hoped to cut 12 800 jobs from the combined workforce, with another 10 800 positions likely to be transferred to lower-cost locations. The two companies would then have a joint workforce of 217 000 worldwide, including the 62 400 staff who work for Barclays in the UK.

Having discussed scale economies, we begin putting flesh on the bare-bones theory of supply we developed in Chapter 6. Despite the growing importance of scale economies, we begin by discussing the output decision of a firm with a general U-shaped average cost curve.⁷

CASE 7.3

GLOBALIZATION, TECHNICAL CHANGE AND INCOME DISTRIBUTION

Over the last 30 years the world economy has become increasingly global. Few restrictions on trade, better transportation technology and the ease with which information can be exchanged have made it possible to increase global trade. Over the same period, income inequality within and across countries has also increased, making the rich richer and the poor poorer. Is globalization to blame?

Recent research undertaken by the International Monetary Fund (IMF) has tried to explain the main determinants of rising income inequality. The data covers 143 countries during 1980–2006. Among the determinants considered are trade globalization (measured as the sum of goods and services each country exchanges with others), financial globalization (measured as the sum of all financial assets moving from and to each country) and technical change (measured as the share of information and communications technology capital in the total capital stock of each country).

The main results are that globalization (trade and financial) can only explain up to 10 per cent of the overall increase in income inequality, while technical change can explain up to 45 per cent. Technical advances favour highly skilled workers, raising their income in comparison to

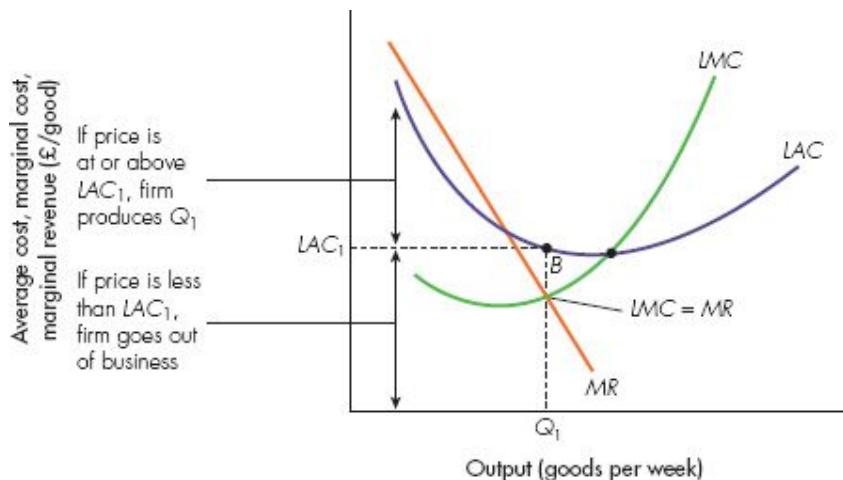
lower skilled workers. This in turn increases income inequality within countries.

In advanced economies the role of globalization in explaining income inequality is actually found to be larger than in developing countries. Globalization does have an effect on income inequality but it is not as relevant as common perceptions suggest. Technical change appears to be the main driving force behind the rise in income inequality in the last few decades.

Source: IMF, ‘Globalization and inequality’, *World Economic Outlook*, 2007.

7.8 The firm’s long-run output decision

Figure 7.11 shows smooth LAC and LMC curves for a firm not restricted to producing integer units of output. It also shows the marginal revenue MR curve. From Chapter 6 we know that the output of maximum profit, or minimum loss, is at B , the output at which marginal revenue equals marginal cost. The firm then checks whether it makes profits or losses at this output. It should not stay in business if it makes losses for ever.



In the long run the firm chooses its output level at the point B here MR is equal to LMC . It has then to check whether it is making losses at that output level Q_1 . If price is equal to or more than LAC_1 , the long-run average cost corresponding to output Q_1 , the firm is not making losses and stays in business. If the price is less than LAC_1 , the firm’s long-run output decision should be zero – it closes down permanently.

Figure 7.11 The firm’s long-run output decision

Total profit is average profit per unit of output, multiplied by output. Total profit is positive only if average profit is positive. Average profit is average revenue minus average cost. But average revenue is simply the price for which each output unit is sold. Hence *if long-run average costs at B exceed the price for which the output Q_1 is sold*, the firm makes losses in the long run and should close down. If, at this output, price equals LAC , the firm just breaks even. If price exceeds LAC at this output, the firm makes long-run profits and happily remains in business.

First, we use the *marginal condition* ($LMC = MR$) to find the best output provided the firm stays in business. Then, we use the *average condition* (comparing LAC at this output with the price or average revenue received) to see if the best positive output yields a profit or a loss.

This is summarized in Table 7.11.

Table 7.11 The firm's output decision in the long run

	Marginal condition	Check whether to produce
Long run	which $MR = LMC$ Choose the output at	Produce this output if $p > LAC$. Otherwise, produce zero.

7.9 The relationship between short-run and long-run average costs

There is a close relationship between short-run and long-run average costs. This relationship is shown in Figure 7.12.

Suppose that in the short run the fixed input of production of a firm is the number of plants (or plant size). Suppose that the firm produces using one plant in the short run. The short-run average cost of the firm is $SATC_1$. Now suppose instead that the firm produces using two plants in the short run. Its short-run average cost will be different and it will be $SATC_2$, and so on. In the long run the plant size will be variable and so the firm can choose the plant size that minimizes the costs.

By definition, the LAC curve shows the least-cost way to make each output when all factors can be varied. A is the least-cost way to make output Q_1 in the short run. B is the least-cost way to make an output Q_2 . It must be more costly to make Q_2 using the wrong quantity of plant, for example the quantity corresponding to point E . For the plant size at A , $SATC_1$ shows the cost of

producing each output, including Q_2 . Hence $SATC_1$ must lie above LAC at every point except A , the output level for which this plant size is best.

This argument can be repeated for other plant sizes. Hence $SATC_3$ and $SATC_4$, reflecting plant sizes at C and D , respectively, must lie above LAC except at points C and D themselves. In the long run the firm can vary all its factors and can generally produce a particular output more cheaply than in the short run, when it is stuck with the quantities of fixed factors it was using previously. A firm currently suffering losses because demand has fallen may make future profits once it has had time to build a plant more suitable for its new output.

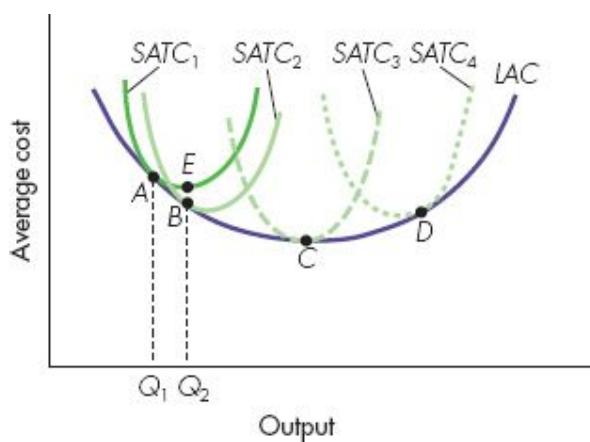


Figure 7.12 The long-run average cost curve LAC

Suppose the plant size is fixed in the short run. For each plant size we obtain a particular $SATC$ curve. But in the long run even plant size is variable. To construct the LAC curve we select at each output the plant size which gives the lowest $SATC$ at this output. Thus points such as A , B , C and D lie on the LAC curve. Notice the LAC curve does not pass through the lowest point on each $SATC$ curve. Thus the LAC curve shows the minimum average cost way to produce a given output when all factors can be varied, not the minimum average cost at which a given plant can produce.

Summary

- This chapter discusses short-run and long-run decisions, based on the corresponding cost curves. In the **long run**, a firm can fully adjust all its inputs. In the **short run**, some inputs are fixed. The length of the short run varies from industry to industry.
- The **production function** shows the maximum output that can be produced using given quantities of inputs. The inputs are machines, raw

materials, labour and any other factors of production. The production function summarizes the technical possibilities faced by a firm.

- In the short run, the firm cannot adjust some of its inputs. But it still has to pay for them. It has short-run fixed costs (SFC) of production. The cost of using the variable factors is short-run variable cost (SVC). Short-run total cost (STC) is equal to SFC plus SVC .
- The **total cost curve** is derived from the production function, for given wages and rental rates of factors of production.
- Short-run average total cost ($SATC$) is equal to short-run total cost (STC) divided by output. **$SATC$ is equal to short-run average fixed cost ($SAFC$) plus short-run average variable cost ($SAVC$)**. The $SATC$ curve is U-shaped. The falling part of the U results both from declining $SAFC$ as the fixed costs are spread over more units of output and from declining $SAVC$ at low levels of output. The $SATC$ continues to fall after $SAVC$ begins to increase, but eventually increasing $SAVC$ outweighs declining $SAFC$ and the $SATC$ curve slopes up.
- The **short-run marginal cost curve** (SMC) reflects the marginal product of the variable factor, holding other factors fixed. Usually we think of labour as variable but capital as fixed in the short run. When very little labour is used, the plant is too big for labour to produce much. Increasing labour input leads to large rises in output and SMC falls. Once machinery is fully manned, extra workers add progressively less to output. SMC begins to rise.
- The SMC curve cuts both the $SATC$ and $SAVC$ curves at their minimum points.
- In the short run, the firm supplies the output at which SMC is equal to MR , provided price is not less than short-run average variable cost. In the short run, the firm is willing to produce at a loss provided it is recovering at least part of its fixed costs.
- The **long-run total cost curve** is obtained by finding, for each output, the least-cost method of production when all inputs can be varied. If the relative price of using a factor of production rises, the firm substitutes away from that factor in its choice of production techniques.

- **Average cost** is total cost divided by output. The **long-run average cost curve** (*LAC*) is derived from the long-run total cost curve.
- ***LAC* is typically U-shaped.** As output rises, at first average costs fall because of indivisibilities in production, the benefits of specialization and engineering advantages of large scale. There are increasing returns to scale on the falling part of the U. The rising part of the U reflects diseconomies of scale.
- Much of manufacturing has **economies of scale**. For some industries, particularly personal services, economies of scale run out at quite low output levels.
- When **marginal cost** is below average cost, average cost is falling. When marginal cost is above average cost, average cost is rising. Average and marginal cost are equal only at the lowest point on the average cost curve.
- In the long run, the firm supplies the output at which **long-run marginal cost** (*LMC*) equals *MR* provided price is not less than the level of long-run average cost at that level of output. If price is less than long-run average cost, the firm goes out of business.
- The *LAC* curve is always below the *SATC* curve, except at the point where the two coincide. This implies that a firm is certain to have higher profits in the long run than in the short run if it is currently producing with a plant size that is not best from the viewpoint of the long run.

Review questions



EASY

- 1 (a) What information does the production function provide? (b) Explain why the production function does not provide enough information for anyone actually to run a firm.
- 2 (a) Calculate the marginal and average costs for each level of output from the following total cost data. (b) Show how marginal and average costs are related. (c) Are these short-run or long-run cost curves? Explain how you can tell.

Output	0	1	2	3	4	5	6	7	8	9
TC (£)	12	27	40	51	60	70	80	91	104	120

- 3 (a) Explain why it might make sense for a firm to produce goods that it can only sell at a loss. (b) Can it keep on doing this forever? Explain.
- 4 **Common fallacies** Why are these statements wrong? (a) Firms making losses should quit at once. (b) Big firms can always produce more cheaply than smaller firms can. (c) Small-scale production is always better.
- 5 The table below shows how output changes as inputs change for three different output levels. The wage rate is £5 and the rental rate of capital is £2.

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Capital input	4	2	7	4	11	8
Labour input	5	6	10	12	15	16
Output	4	4	8	8	12	12

- a. For each output level in the above table, which technique of production is more capital intensive?
- b. Refer to columns 2, 3 and 6. Does the firm switch towards or away from more capital-intensive techniques as output rises?
- 6 For each of the following cases explain how long you think the short run is:
 a) a power station; (b) a supermarket; (c) a small grocery retail business. In explaining your answer, specify any assumptions you need to make. For each case, do you expect the law of diminishing marginal returns to hold?
- 7 Suppose the rental rate of capital in the table below is £3 and the wage rate is £5. (a) Suppose the firm is currently using the production techniques shown in columns 1, 3 and 5 for output levels 4, 8 and 12, respectively. Would the firm switch to columns 2, 4 or 6, respectively, for any levels of output? (b) How do the firm's total and average costs change when the rental rate of capital rises?

MEDIUM

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Capital input	4	2	7	4	11	8
Labour input	5	6	10	12	15	16
Output	4	4	8	8	12	12
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Capital input	4	2	7	4	11	8
Labour input	5	6	10	12	15	16
Output	4	4	8	8	12	12

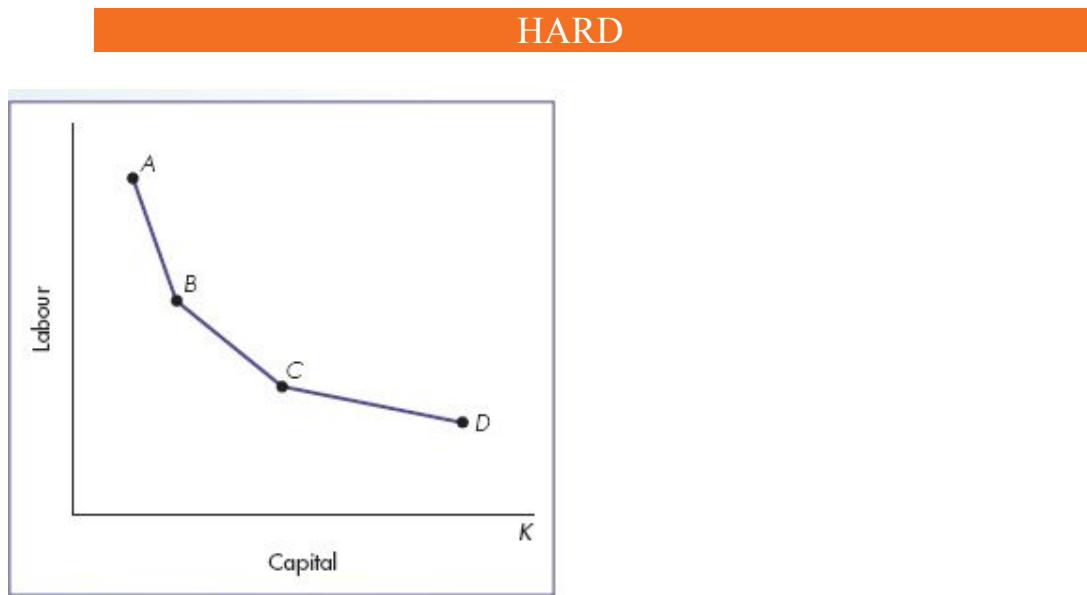
- 8 The marginal cost of supplying another unit of output of an electronic product via the Internet is almost zero. If long-run equilibrium means price equals marginal cost, all Internet firms will go bust. Can you resolve this situation?
- 9 What kind of returns to scale do the following production functions display?

(a) $Q = \sqrt{KL}$ (b) $Q = K^{0.3} L^{0.2}$ (c) $Q = K + L$

- 10 (a) What are economies of scale and why might they exist? (b) The table below shows how output changes as inputs change. The wage rate is £5 and the rental rate of capital is £2. Calculate the lowest-cost method of making 4, 8 and 12 units of output. (c) Are there increasing, constant or decreasing returns to scale between those outputs? Which applies where?

Capital input	4	2	7	4	11	8
Labour input	5	6	10	12	15	16
Output	4	4	8	8	12	12

- 11 The following graph shows the isoquant for a firm. What do points A, B, C and D on the isoquant represent? Which is the most labour intensive and the most capital intensive technique of production on this isoquant?



- 12 Suppose that firm A has the following short-run production function $Q = K_c \sqrt{L}$, where K denotes capital and L labour. Suppose that the level of capital is fixed at $k_0 = 10$. The total cost of firm A in the short run is $STC = 10wL$ where w is the wage paid to each worker. Assume that the wage is £20. Using the production function, show how the short-run total cost depends on the quantity produced Q . Plot the short-run total cost on a graph, where you put Q on the horizontal axis.

|3 The following table shows data about quantity produced and total cost of production in the long run for a given firm: Find the long-run marginal cost and the long-run average cost faced by the firm. On a graph, plot the *LMC* and *LAC* curves. Explain why the *LMC* curve cuts the *LAC* curve from below.

Q	LTC
1	102
2	112
3	136
4	180
5	250
6	352
7	492
8	676
9	910
10	1200

|4 **Essay question** We choose between couriers such as DHL and Federal Express based on the quality, convenience and reliability of service that they offer, not just on the price that they quote. Once we recognize that service matters, the inevitability of scale economies is greatly reduced. Even Amazon has to organize the distribution of the products it sells. Do you agree?

Appendix

Isoquants and the choice of production technique in the long run

The choice of technique in the long run can be examined using techniques similar to the indifference curve – budget line approach used to study consumer choice in Chapter 5. Figure 7.A1 plots input quantities of capital *K* and labour *L*. Points *A*, *B*, *C* and *D* show the *minimum* input quantities needed to make 1 unit of output using each of four different techniques. An isoquant shows minimum combinations of inputs to make a given output. Different points on an isoquant reflect different production techniques. Technique *A* is the most labour intensive, requiring L_A units of labour and K_A units of capital to make 1 unit of output. Technique *D* is the most capital intensive.

Connecting A , B , C and D yields an **isoquant** (iso = the same, quant = quantity).

Figure 7.A1 shows four techniques but we can imagine that there are others.

An **isoquant** shows minimum combinations of inputs to make a given output. Different points on an isoquant reflect different production techniques.

In practice, the isoquants give us a graphical description of the technology available to a firm. Technology is summarized by the production function. Therefore we can derive isoquants directly from a given production function. Suppose that a firm is producing using the following technology:

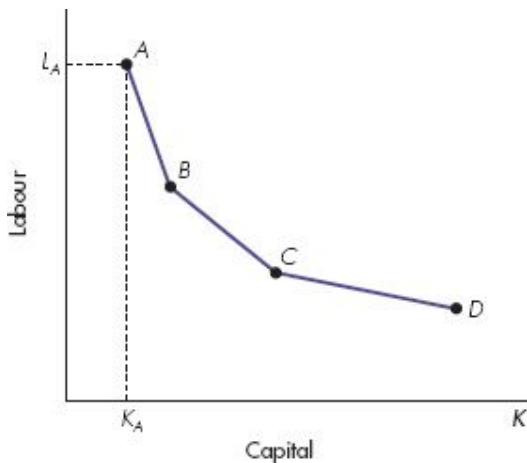
$$Q = KL \quad (\text{A1})$$

For example, if our firm uses 2 units of labour ($L = 2$) and 5 units of capital ($K = 5$) it can produce 10 units of output. What is an isoquant for our firm if it wants to produce 10 units of output? By setting $[Q=10]$ into equation (A1), we get $[10=KL]$ or written differently:

$$L = \frac{10}{K} \quad (\text{A2})$$

This is the equation of an isoquant when output is equal to 10. It gives us all the combinations of capital and labour needed to produce $Q = 10$. For example, $L = 2$ and $K = 5$ is such a combination, but so too is $L = 1$ and $K = 10$, $L = 5$ and $K = 2$ and so on. If we plot (A2) we have a particular isoquant, like the one displayed in Figure 7.A1. We can do the same for different levels of output and so we can have different isoquants, one for each level of output. In practice, we can create a map of isoquants in pretty much the same way as we did for indifference curves in Chapter 5.

In general, given a production function $F(K, L)$, the equation of an isoquant is $F(K, L) = Q_0$ where Q_0 is a specified value for the output level. Figure 7.A2 shows a map of smooth isoquants for a given production function. Isoquant I corresponds to a particular output. Each point on isoquant I reflects a different technique, from very capital intensive to very labour intensive.

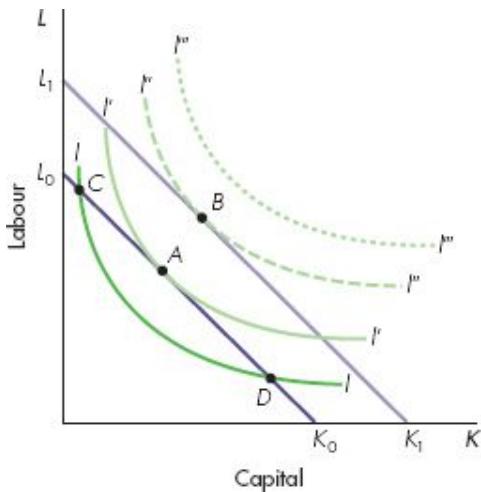


Points A, B, C and D show different input combinations required to produce 1 unit of output. By connecting them we obtain an isoquant that shows the different input combinations which can produce a particular level of output.

Figure 7.A1 An isoquant

Higher isoquants, such as I'' , show higher output levels since more inputs are required.

Three properties of isoquants are important. First, they cannot cross. Each isoquant refers to a different output. Second, each isoquant slopes down. To make a given output, a technique can use more capital only if it uses less labour and vice versa. Hence isoquants must slope down. Third, each isoquant becomes flatter as we move along it to the right, as Figure 7.A2 shows. Moving down a given isoquant, it takes more and more extra capital input to make equally successive reductions in the labour input required to produce a given output. This property is known as the *diminishing marginal rate of technical substitution*. The *marginal rate of technical substitution (MRTS)* tells us the rate at which a firm can substitute capital with labour holding output constant. The marginal rate of technical substitution between labour and capital is defined as:



Each isoquant such as l shows a particular output level. Higher isoquants such as l'' show higher output levels. Straight lines such as L_0K_0 are isocost lines showing different input combinations having the same total cost. The slope of an isocost line depends only on relative factor prices. A higher isocost line such as L_1K_1 implies a larger total cost. To produce a given output, such as that corresponding to the isoquant l' , the firm chooses the point of tangency of that isoquant to the lowest possible isocost line. Thus point A is the cost-minimizing way to produce the output level on l' and point B the cost-minimizing way to produce the output level on l'' .

Figure 7.A2 Cost minimization

$$MRTS = \frac{\Delta L}{\Delta K} \quad (\text{A3})$$

This tells us that if we decrease the amount of capital used in production by ΔK units, then we need to increase the amount of labour employed by ΔK units to keep the amount of output produced unchanged.

If the change in capital and the change in labour are small enough (close to zero), we can use the idea of a derivative and write: $MRTS = dL/dK$. In this case, the $MRTS$ is just the slope of an isoquant at a given point.⁸

The isoquant tells us how much output a given technique can produce. However, to choose the right technique we need to know how much a given technique costs. This information is given by the *isocost* line. In Figure 7.A2 the line L_0K_0 is an isocost line. It shows different input combinations with the *same* total cost. For a given cost, the firm can use more units of capital only if it uses fewer units of labour.

The isocost line is derived from the cost function. Suppose that a firm faces the following total cost function by employing labour and capital:

$$C = wL + rK \quad (\text{A4})$$

where w is the price the firm must pay to employ one unit of labour, that is, the wage; r is the price of capital, that is, the rental rate; and C denotes the total cost.

To find the isocost line we just fix the total cost C to a given value, say C_0 , and then find the combination of labour and capital, given their prices, that results in such a cost. By substituting $C = C_0$, into (A4) and solving for L , we find:

$$L = \frac{C_0}{w} - \frac{r}{w}K \quad (\text{A5})$$

Equation(A5) is the typical equation of an isocost line, such as line L_0K_0 in Figure 7.A2. For example, suppose $C_0 = 100$, $w = 1$ and $r = 5$, then (A5) becomes $L = 100 - 5K$. A combination of capital and labour that costs 100 at those prices is then $K = 10$ and $L = 50$. Another one is $K = 15$ and $L = 25$ and so on.

The main features of the isocost line are the following: first, the slope of the isocost line reflects the relative price of the two factors of production (r/w). Consider Figure 7.A2 and the isocost line L_0K_0 . Beginning at K_0 , where all the firm's money is spent on capital, the firm can trade off 1 unit of capital for more units of labour the cheaper the wage rate relative to the rental cost of capital. Second, facing given factor prices, by raising spending a firm can have more capital and more labour. A higher isocost line parallel to L_0K_0 shows a higher spending on inputs. Along the isocost line L_1K_1 the firm spends more on inputs than along the isocost line L_0K_0 .

To minimize the cost of making a given output, a firm chooses the point of tangency of that isoquant to the lowest possible isocost line. At this point, the (negative) slope of the isocost line equals the (negative) slope of the isoquant. We know that the slope of the isocost line is $-r/w$. The slope of an isoquant is the marginal rate of technical substitution. If we increase the amount of capital by ΔK units, the firm gains $(MP_K)\Delta K$ units of output, where MP_K is the marginal product of capital. By decreasing the amount of labour by ΔL units, the firm gives up $(MP_L)\Delta L$ units of output, where MP_L is the marginal product of labour. Since output is constant along an isoquant, those change must leave unchanged the amount of output and therefore it must be true that:

$$MP_K\Delta K + MP_L\Delta L = 0$$

We can rearrange that equation in the following way:

$$\frac{\Delta L}{\Delta K} = -\frac{MP_K}{MP_L} \quad (\text{A6})$$

The left-hand side is just the $MRTS$.

The isoquant's slope [$2MP_K/MP_L$] tells us by how much labour is changed to keep output constant when capital is changed by ΔK units. Hence the tangency condition in Figure 7.A2 implies

$$\text{Slope of isocost line} = -r/w = -MP_K/MP_L = \text{slope of isoquant} \quad (\text{A7})$$

Point A in Figure 7.A2 is the least-cost way to make the output shown by isoquant I' . We can repeat this analysis for every other isoquant showing different outputs. By doing that, we can derive the total long-run cost curve faced by a firm, as shown in Figure 7.A3.

In Figure 7.A3 we have three possible isoquants corresponding to three different levels of output, with $Q_1 < Q_2 < Q_3$. We also have three isocosts corresponding to different levels of cost, with $C_1 < C_2 < C_3$. The corresponding cost-minimizing choices of input for each level of output are points A , B and C . In Panel 2 we translate the information from Panel 1 into a graph, where we put total costs (TC) on the vertical axis and level of output on the horizontal axis. By producing Q_1 the total cost faced by the firm is C_1 . This is point A in Panel 2. By producing Q_2 the firm will pay a total cost of C_2 , which is point B , and so on. By connecting all those points, we draw the long-run total cost curve (LTC) of our firm. Notice that each point of the long-run total cost curve represents the minimum cost a firm has to pay to produce the quantity at that point.

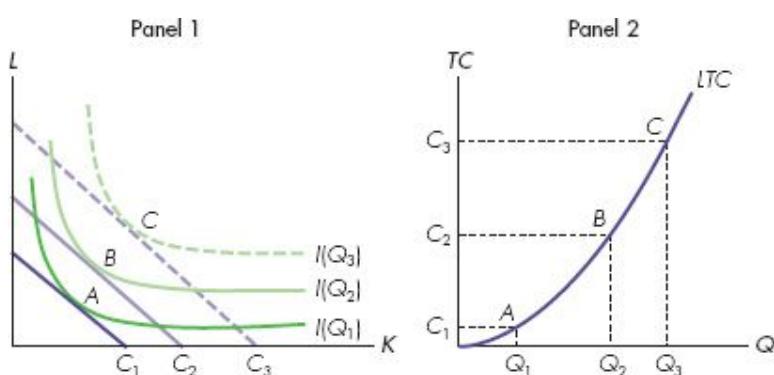
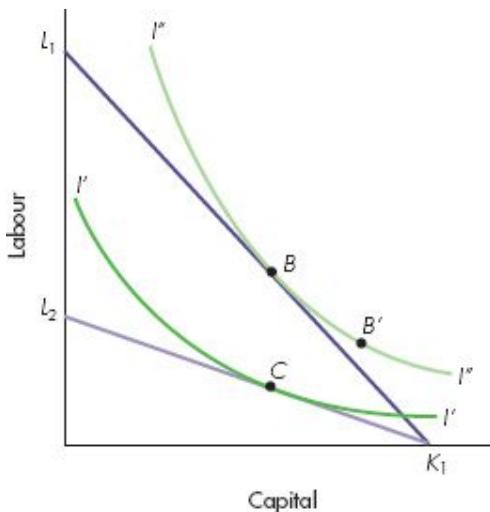


Figure 7.A3 Cost minimization and long-run total cost curve



At the old factor prices the firm's output level corresponds to the isoquant I'' . All isocost lines have the same slope as $L_1 K_1$ and the firm produces its given output most cheaply by choosing the point B where the isoquant is tangent to the lowest possible isocost line $L_1 K_1$. A wage increase makes all isocost lines flatter, parallel to $L_2 K_1$. Each unit of capital sacrificed now allows the purchase of less additional labour. The wage increase has a pure substitution effect from B to B' where the original isoquant I'' has the same slope as the new isocost lines. Firms substitute capital for labour. But with higher marginal costs at each output level, the firm's profit-maximizing output is reduced, say to the level corresponding to the isoquant I' . On this isoquant, costs are minimized by producing at C to reach the lowest possible isocost line $L_2 K_1$ at the new factor prices. The move from B' to C is the pure effect induced by the shift in the firm's marginal cost curve for its output.

Figure 7.A4 The effect of a wage increase

Finally, we show the effect of a rise in the price of one factor on the cost-minimizing choice of techniques and the profit-maximizing level of output. In Figure 7.A4 the cost-minimizing choice of inputs is at point B , where condition (A7) holds since the isocost line $L_1 K_1$ is tangent to the isoquant I'' . Given the cost-minimizing choice of inputs, the long-run profit-maximizing output is the output associated with isoquant I'' .

Suppose the wage rate rises. Each isocost line then becomes less steep. Sacrificing a unit of labour allows more extra capital at any given total cost. The new isocost line is $L_2 K_1$ and the new cost-minimizing choice of inputs is at point C on a lower isoquant. At the original output on isoquant I'' , the increase in the wage leads to a pure substitution effect from B to B' ; the point on the old isoquant is tangent to an isocost line with the new flatter slope. The firm substitutes the now relatively more expensive labour with capital. But a higher wage rate also shifts up the total cost curve and the marginal cost curve for output. Profit-maximizing output falls.

The lower isoquant I' shows the new profit-maximizing output. The move from B' to C is the pure output effect of a higher wage rate. The actual move from B to C can be decomposed into a substitution effect from B to B' and an output effect from B' to C . Both effects reduce the quantity of labour demanded and the profit maximizing level of output.

- 1 Economists use diminishing returns to describe the addition of one variable factor to other fixed factors in the short run, but decreasing returns to describe diseconomies of scale when all factors are freely varied in the long run.
- 2 The Appendix to this chapter analyses the concept of cost function in more detail.
- 3 The same idea was expressed more formally in Maths 6.2.
- 4 This problem is analysed in more detail in the Appendix to this chapter.
- 5 Since output, revenue and cost are all flows, these should be measured per week or per year. We omit time units for brevity but do not forget they are flows not stocks!
- 6 See F. M. Scherer and D. Ross, Industrial market structure and economic performance, 3rd edn (Houghton Mifflin, 1990).
- 7 In Chapter 8, Section 8.10, we discuss the case when a firm faces significant economies of scale.
- 8 If we plot an isoquant on a graph where capital is on the vertical axis and labour is on the horizontal axis (that is, the opposite of Figures 7.A1 and 7.A2), then the MRTS is normally computed as DK/DL .

CHAPTER 8

Perfect competition and pure monopoly

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 define perfect competition
- 2 describe why a perfectly competitive firm equates marginal cost and price
- 3 understand how profits and losses lead to entry and exit
- 4 draw the industry supply curve
- 5 understand comparative static analysis of a competitive industry
- 6 define pure monopoly
- 7 understand why a monopolist's output equates MC and MR
- 8 recognize how output compares under monopoly and perfect competition
- 9 describe how price discrimination affects a monopolist's output and profits

An industry is the set of all firms making the same product. The output of an industry is the sum of the outputs of its firms. Yet different industries have very different numbers of firms. Eurostar is the only supplier of train journeys from London to Paris. In contrast, the UK has 150 000 farms and 20 000 grocers.

In a **perfectly competitive** market, both buyers and sellers believe that their own actions have no effect on the market price. In contrast, a **monopolist**, the only seller or potential seller in the industry, sets the price.

Why do some industries have many firms but others only one? Chapter 9 develops a general theory of market structure,¹ showing how demand and cost conditions together determine the number of firms and their behaviour.

First, it is useful to establish two benchmark cases; that is, extremes between which all other types of market structure must lie. These limiting cases are **perfect competition** and **monopoly**.

We focus on how the number of sellers affects the behaviour of sellers. Buyers are in the background. We simply assume there are many buyers whose individual downward-sloping demand curves can be aggregated into the market demand curve. Thus, we assume that the demand side of the market is competitive but contrast the different cases on the supply side.²

Perfect competition means that each firm or household, recognizing that its quantities supplied or demanded are trivial relative to the whole market, assumes its actions have no effect on the market price. This assumption was built into our model of consumer choice in Chapter 5. Each consumer's budget line took market prices as given, unaffected by the quantities then chosen. Changes in *market* conditions, applying to all firms and consumers, change the equilibrium price and hence individual quantities demanded, but each consumer neglects any effect on market price created by his own actions.

This concept of competition, which we now extend to firms and supply, differs from everyday usage. Ford and Volkswagen are fighting each other for the European car market but an economist would not call them perfectly competitive. Each has such a big share of the market that changes in the quantity supplied by either firm affect the market price. Ford and Volkswagen each take account of this in deciding how much to supply. They are not *price-takers*. Only under perfect competition do individuals make decisions that treat the price as independent of their own actions.

8.1 Perfect competition

In the US there are approximately 400 000 farms producing corn. The corn produced on one farm is probably the same as the corn produced on all other farms. Therefore, all farms produce the same good. Given the number of farms, it is unlikely that the production of one farm affects the production of all other farms and, thus, the market price of corn. Indeed the price of corn is determined at the Chicago Board of Trade (CBOT), a formal exchange market (like a stock market for commodities) where buyers and sellers post their orders. Therefore we can say that each corn farm in the US takes the price of corn as given when they decide how much corn to produce. The corn industry in the US is an example of a market that resembles a perfectly competitive market.

In a perfectly competitive industry an individual's action does not affect the price. Therefore in such an industry there must be many sellers. Each firm in a perfectly competitive industry faces a horizontal demand curve, as shown in Figure 8.1. However much the firm sells, it gets the market price. If it charges a price above P_0 , it will not sell any output: buyers will go to other firms whose product is just as good. Since the firm can sell as much as it wants at P_0 , it will not charge less than P_0 . The individual firm's demand curve is DD .

A *horizontal* demand curve, along which the price is fixed, is the key feature of a perfectly competitive firm. To be a plausible description of the demand curve facing the firm, the industry must have four attributes. First, there must be many firms, each trivial relative to the entire industry. Second, the product must be standardized. This means that all firms produce *homogeneous* or identical goods. Even if the car industry had many firms it would not be a competitive industry. A Ford Mondeo is not a perfect substitute for a Vauxhall Astra. The more imperfect they are as substitutes, the more it makes sense to view Ford as the sole supplier of Mondeos and Vauxhall as the sole supplier of Astras. Each producer ceases to be trivial relative to the relevant market and cannot act

as a price-taker. In a perfectly competitive industry, all firms must be making the same product, *for which they all charge the same price*.

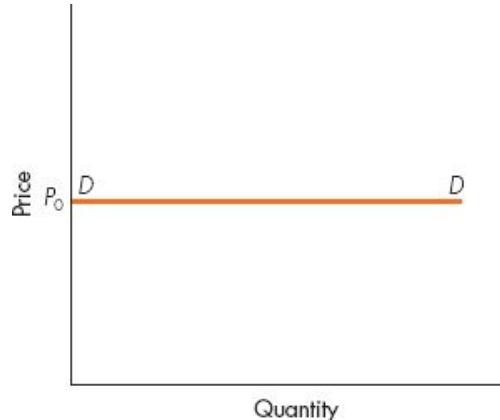


Figure 8.1 The competitive firm's demand curve

A competitive firm can sell as much as it wants at the market price P_0 . Its demand curve DD is horizontal at this price.

Third, even if all firms in an industry made identical goods, each firm may have some discretion over the price it charges if buyers have imperfect information about the quality or characteristics of products. To rule this out in a competitive industry, we must assume that buyers have almost *perfect information* about the products being sold. They know that the products of different firms in a competitive industry really are identical.

The fourth crucial characteristic of a perfectly competitive industry is *free entry and exit*. Even if existing firms could organize themselves to restrict total supply and drive up the market price, the consequent rise in revenues and profits would simply attract new firms into the industry, thereby increasing total supply again and driving the price back down.

Conversely, as we shall shortly see, when firms in a competitive industry are losing money, some firms will close down and, by reducing the number of firms remaining in the industry, reduce the total supply and drive the price up, thereby allowing the remaining firms to survive.

To sum up, each firm in a competitive industry faces a horizontal demand curve at the going market price. To be a plausible description of the demand conditions facing a firm, the industry must have: (1) many firms, each trivial relative to the industry; (2) a homogeneous product, so that buyers would switch between firms if their prices differed; (3) perfect customer information about product quality, so that buyers know that the products of different firms really are the same; and (4) free entry and exit, to remove any incentive for existing firms to collude.

CONCEPT 8.1

WHY DO WE NEED TO STUDY PERFECTLY COMPETITIVE MARKETS?

A perfectly competitive market is characterized by many firms producing an identical product and so each firm is a price-taker, firms have freedom of entry and exit, and buyers are perfectly informed about the product sold in the market. All those assumptions rarely hold together in a given market, and therefore we rarely see a perfectly competitive market in reality.

Why do we need to study something that may not exist in the real world?

The answer is that we need to study perfectly competitive markets, for two main reasons. First, a perfectly competitive market may be considered a good approximation for many markets. For example, markets such as those for agricultural products, the stock market, the housing market and so on, can be reasonably well described by using the theory we are going to develop in this chapter (together with the one developed in Chapter 3).

8.2 A perfectly competitive firm's supply decision

Chapter 7 developed a general theory of supply. The firm uses the marginal condition ($MC = MR$) to find the best positive output. Then it uses the average condition to check whether the price for which this output is sold covers average cost.

This general theory must hold for the special case of perfectly competitive firms. The *special feature of perfect competition is the relationship between marginal revenue and price*. A competitive firm faces a horizontal demand curve. Making and selling extra output does *not* bid down the price for which existing output is sold. The extra revenue from selling an extra unit is simply the price received. A perfectly competitive firm's marginal revenue is its output price:

The **short-run supply curve** is the SMC curve above the point at which the SMC curve crosses the lowest point on the SAVC curve.

$$MR = P$$

A firm's short-run supply curve

Figure 8.2 shows again the short-run cost curves – marginal cost SMC , average total cost $SATC$ and average variable cost $SAVC$ – from Chapter 7. Any firm chooses the output at which marginal cost equals marginal revenue. Equation (1) means that a perfectly competitive firm chooses the output at which

$$SMC = MR = P$$

Suppose the firm faces a horizontal demand curve at the price P_4 in Figure 8.2. From equation (2), the firm chooses the output Q_4 to reach point D , at which price equals marginal cost.

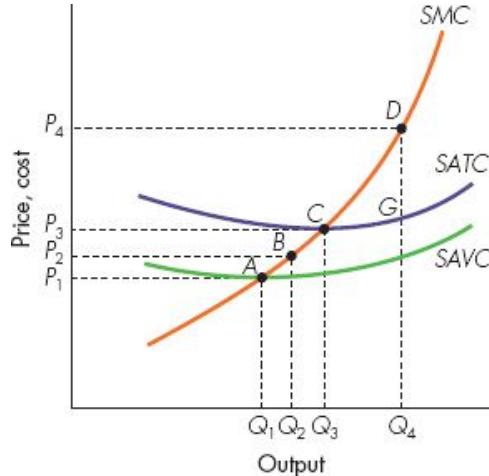


Figure 8.2 Short-run supply decisions of the perfectly competitive firm

The perfectly competitive firm produces at that level of output at which price is equal to marginal cost, provided it makes more profit by producing some output than none at all. The firm's short-run supply curve is the SMC curve above the point A , the shutdown point below which the firm cannot cover average variable costs $SAVC$ in the short run.

Next, the firm checks whether it would rather shut down in the short run. It shuts down if the price P_4 fails to cover short-run variable cost at this output. In Figure 8.2 P_4 exceeds $SAVC$ at the output Q_4 . The firm supplies Q_4 and makes profits. Point D lies above point G , the short-run average total cost (including overheads) of producing Q_4 . Hence profits are the rectangle obtained by multiplying the vertical distance DG (average profit per unit produced) by the horizontal distance OQ_4 (number of units produced).

In the short run, the firm supplies positive output for any price above P_1 . At a price P_2 , the firm makes Q_2 , the output at which price equals marginal cost. Any price below P_1 is below the minimum point on the $SAVC$ curve. The firm cannot find an output at which price covers $SAVC$. Between points A and C , the firm is making short-run losses but recouping some of its overheads. At any price above P_3 , at which the SMC curve crosses the lowest point on the $SATC$ curve, the firm is making short-run profits.

Remember that these are economic or supernormal profits. A firm's long-run supply curve, relating output supplied to price in the long run, is that part of its LMC curve above its LAC curve, after allowing for the economic costs, including the opportunity costs of the owner's financial capital and work effort, summarized in the $SAVC$ and $SATC$ curves.

A firm's long-run supply curve

A firm's long-run supply curve, relating output supplied to price in the long run, is that part of its LMC curve above its LAC curve.

Figure 8.3 shows the firm's average and marginal costs in the long run. The long-run marginal cost curve LMC is flatter than the SMC curve since the firm can adjust all inputs in the long run.

Facing a price P_4 , the firm chooses the long-run output Q_4 at point D , then checks whether it is better to shut down than to produce this output. In the long run, shutting down means leaving the industry altogether. The firm exits the industry if price fails to cover long-run average cost LAC at the best positive output. At the price P_2 , the best positive output is at point B in Figure 8.3, but the firm makes a loss and should exit the industry in the long run. At any price below P_3 , the firm exits the industry. At the price P_3 , the firm produces Q_3 and just breaks even after paying all its economic costs. It makes only **normal profits**.

When economic profits are zero the firm makes **normal profits**. Its accounting profits just cover the opportunity cost of the owner's money and time.

Entry and exit

Entry is when new firms join an industry.

Exit is when existing firms leave.

The price P_3 , corresponding to the lowest point on the LAC curve in Figure 8.3, is the entry or exit price. Firms make only normal profits. There is no incentive to enter or leave the industry. The resources tied up in the firm are earning just as much as their opportunity costs; that is, what they could earn elsewhere. Any price below P_3 induces the firm to exit the industry in the long run. P_3 is the minimum price required to keep the firm in the industry.

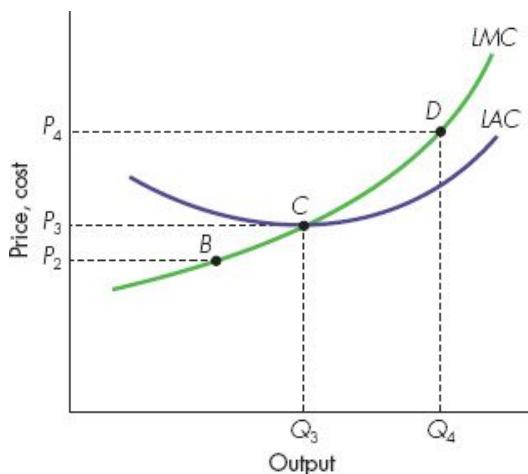


Figure 8.3 Long-run supply decisions of the perfectly competitive firm

The perfectly competitive firm produces at that level of output at which P is equal to marginal cost, provided it makes more profit by producing some output than none at all. It therefore chooses points on the LMC curve. At any price above P_3 the firm makes profits because price is above long-run average cost (LAC). At any price below P_3 , such as P_2 , the firm makes losses because price is below long-run average cost. It therefore will not produce any output at prices below P_3 . The long-run supply curve is the LMC curve above point C.

We can also interpret Figure 8.3 as the decision facing a potential entrant to the industry. The cost curves now describe the post-entry costs. P_3 is the price at which entry becomes attractive. Any price above P_3 yields supernormal profits and encourages entry of new firms.

The marginal firm

In the short run, the number of firms in a perfectly competitive market is fixed. In the long run, the number of firms is determined by entry and exit. If in the long run firms in the market are making supernormal profits, then other firms may find it profitable to enter the market. Assume that all firms in the market and the potential entrants are equal, meaning they have the same cost curves.

Consider Figure 8.4. Suppose that market demand and market supply are such that the market price is P_1 . Given the long-run average cost curve depicted in Figure 8.4, when the price is P_1 , a firm in the market makes supernormal profits. New firms may then enter the market. The main effect of this entry is that more firms will produce in the market and so the market supply will shift to the right. This is represented by the shift from S_1 to S_2 in Figure 8.4. Since the supply has increased, for a given market demand, the market price will decrease. This will reduce the profits that firms in the market can make.

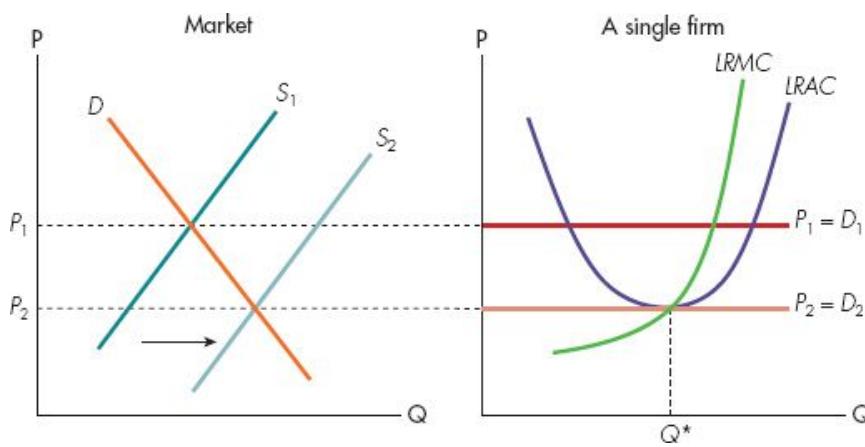


Figure 8.4 Entry to and exit from the industry

When will entry into the market stop? When the last firm to enter makes zero profits. This last firm to enter is called the **marginal firm**. In Figure 8.4, entry stops when the market price is P_2 . That price will be equal to the minimum $LRAC$ for all firms and so each firm will make zero profit. No other firm will find entry profitable since there are no profits to steal. When the market price is above P_2 , entry is profitable. At P_2 , entry stops and the market is in long-run equilibrium.

The **marginal firm** is the last firm to enter the market; it makes zero long-run profits.

Supply decisions of a competitive firm

The **shutdown price** is the price below which the firm cuts its losses by making no output.

Figure 8.5 summarizes the preceding discussion. For each level of fixed factors there is a different SMC curve and short-run supply curve $SRSS$. The long-run supply curve $LRSS$ is flatter than $SRSS$ because extra factor flexibility in the long run makes the LMC curve flatter than the SMC curve. The $SRSS$ curve starts from a lower **shutdown price** because, in the short run, a firm will produce if it can cover average variable costs. In the long run, all costs are variable and must be covered if the firm is to stay in the industry. In either case, a competitive firm's supply curve is the part of its marginal cost curve above the point at which it is better to make no output at all. Table 8.1 sets out this principle.

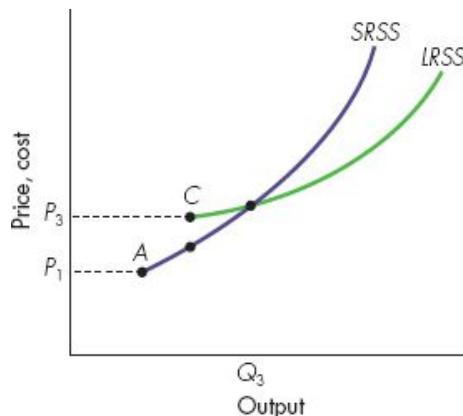


Figure 8.5 Short- and long-run supply curves of the competitive firm

Taken from the two previous figures, the short-run supply curve SRSS is the firm's SMC curve above A and the long-run supply curve LRSS is the firm's LMC curve above C. P_1 is the shutdown price in the short run and P_3 the entry and exit price in the long run. If the firm happens to begin with the stock of fixed factors it would choose at the lowest point on its LAC curve, then C will actually lie on the SRSS curve.

Table 8.1 Supply decisions of a perfectly competitive firm

Marginal condition	Average condition	Short run	Long run
Produce output where $P = MC$	If $P > SAC$, shut down temporarily If $P > LAC$, exit industry		

8.3 Industry supply curves

A competitive industry comprises many firms. In the short run, two things are fixed: the quantity of fixed factors used by each firm and the number of firms in the industry. In the long run, each firm can vary all its factors of production but the number of firms can also change through entry to and exit from the industry.

The short-run industry supply curve

Figure 8.6 adds firms' individual supply curves to ascertain the industry supply curve. At each price, we add the quantities supplied by each firm to find the total quantity supplied at that price.

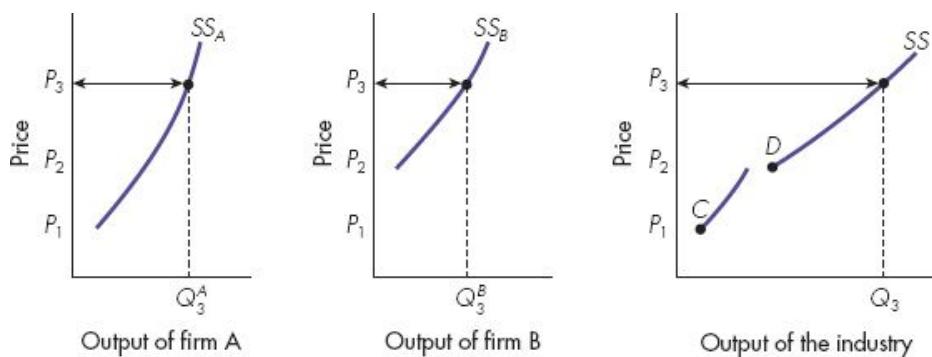


Figure 8.6Deriving the industry supply curve

The industry supply curve SS shows the total quantity supplied at each price by all the firms in the industry. It is obtained by adding at each price the quantity supplied by each firm in the industry. With only two firms, A and B, the figure shows how, at a price such as P_3 , we add Q_3^A and Q_3^B to obtain Q_3 on the industry supply curve. Since firms can have different shutdown prices or entry and exit prices, the industry supply curve can have step jumps at points such as C and D where an extra firm starts production. However, with many firms in the industry, each trivial relative to the industry as a whole, the step jumps in the industry supply curve when another starts production are so small that we can effectively think of the upward-sloping industry supply curve as smooth.

In the short run, the number of firms in the industry is given. Suppose there are two firms, A and B. Each firm's short-run supply curve is the part of its SMC curve above its shutdown price. In Figure 8.6, firm A has a lower shutdown price than firm B. Firm A has a lower $SAVC$ curve. It may have a better location or better technical know-how. Each firm's supply curve is horizontal at the shutdown price. At a lower price, no output is supplied.

At each price, the industry supply Q is the sum of Q_A , the supply of firm A, and Q_B , the supply of firm B. Thus if P_3 is the price, $Q_3 = Q_3^A + Q_3^B$. The industry supply curve is the horizontal sum of the separate supply curves. The industry supply curve is discontinuous at the price P_2 . Between P_1 and P_2 , only the lower-cost firm A is producing. At P_2 , firm B starts to produce as well.

With many firms, each with a different shutdown price, there are many tiny discontinuities as we move up the industry supply curve. Since each firm in a competitive industry is trivial relative to the total, the industry supply curve is effectively smooth.

Comparing short-run and long-run industry supply curves

Figure 8.6 may also be used to derive the long-run industry supply curve. For each firm, the individual supply curve is the part of its LMC curve above its entry and exit price. Unlike the short run, the number of firms in the industry is no longer fixed. Existing firms can leave the industry and new firms can enter. Instead of horizontally aggregating at each price the quantities supplied by the existing firms in the industry, we must horizontally aggregate the quantities supplied by existing firms *and firms that might potentially enter the industry*.

At a price below P_2 in Figure 8.6, firm B is not in the industry in the long run. At prices above P_2 , firm B is in the industry. As the market price rises, total industry supply rises in the long run, not just because each existing firm moves up its long-run supply curve but also because new firms join the industry.

Conversely, at low prices, high-cost firms lose money and leave the industry. Entry and exit in the long run are analogous to shutdown in the short run. In the long run, entry and exit affect the number of producing firms whose output is horizontally aggregated to get the industry supply. In the short run, the number of firms in the industry is given but some are producing while others are temporarily shut down. Again, the industry supply curve is the horizontal sum of those outputs produced at the given market price.

The long-run supply curve is flatter than its short-run counterpart. Each firm can vary its factors more appropriately in the long run and has a flatter supply curve. Moreover, higher prices attract extra firms into the industry. Industry output rises by more than the extra output supplied by the firms already in the industry.

Conversely, if the price falls, firms initially move down their (relatively steep) short-run supply curves. If short-run average variable costs are covered, firms may not reduce output very much. In the long run, each firm reduces output further since all factors of production can now be varied. In addition, some firms exit the industry since they are no longer covering long-run average costs. A price cut reduces industry output by more in the long run than in the short run.

A horizontal long-run industry supply curve

Each firm has a rising LMC curve and thus a rising long-run supply curve. The industry supply curve is a bit flatter. Higher prices do not merely induce existing firms to produce more but also induce new firms to enter. In the extreme case, the industry long-run supply curve is horizontal if all existing firms and potential entrants have *identical cost* curves (Figure 8.7). Below P^* , no firm wants to supply. It takes a price P^* to induce each individual firm to make Q_1 .

At a price P_2 above P^* , each firm makes Q_2 and earns supernormal profits. Point D is above point E. Since potential entrants face the same cost curves, new firms flood into the industry. The industry supply curve is horizontal in the long run at P^* . It is not necessary to bribe existing firms to move up their individual supply curves. Industry output is expanded by the entry of new firms alone. Figure 8.7 shows the long-run industry supply curve $LRSS$, horizontal at the price P^* .

There are two reasons why a rising long-run industry supply curve is much more likely than a horizontal long-run supply curve for a competitive industry. First, it is unlikely that every firm and potential firm in the industry has identical cost curves. Second, even if all firms face the same cost curves, we draw a cost curve for given technology *and* given input prices. Although each small firm affects neither output prices nor input prices, collective expansion of output by all firms may bid up input prices. It then needs a higher output price to induce industry output to rise. In general, the long-run industry supply curve slopes up.

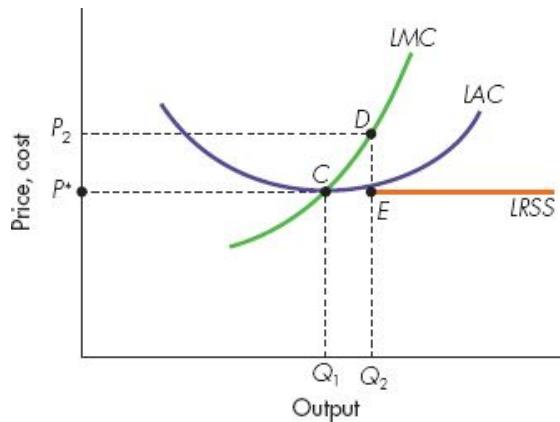


Figure 8.7 The horizontal long-run industry supply curve

When all existing firms and potential entrants have identical costs, industry output can be expanded without offering a price higher than P^* . The long-run industry curve is the horizontal line LRSS at P^* . Industry output can be indefinitely expanded at this price by increasing the number of firms that each produce Q_1 .

8.4 Comparative statics for a competitive industry

Having discussed the industry supply curve, we can now examine how supply and demand interact to determine equilibrium price in the short run and the long run.

We now examine equilibrium in a competitive industry and apply the method of comparative statics. Comparative statics examines how equilibrium changes when demand or cost conditions shift.

Comparative statics examines how equilibrium changes when demand or cost conditions shift.

An increase in costs

Consider a rise in costs, such as a higher input price, that hits all firms in the industry. For simplicity, suppose all firms have the same cost curves and the long-run industry supply curve is horizontal.

In Figure 8.8 the competitive industry faces a downward-sloping demand curve DD . Initially, the long-run supply curve is $LRSS_1$. The market clears at the price P_1^* and the total output Q_1^* . The short-run industry supply curve is $SRSS_1$. The market is in **short-run and long-run equilibrium**.

In **short-run equilibrium**, the price equates the quantity demanded to the total quantity supplied by the given number of firms in the industry when each firm is on its short-run supply curve.

In **long-run equilibrium**, the price equates the quantity demanded to the total quantity supplied by the number of firms in the industry when each firm is on its long-run supply curve and firms can freely enter or exit the industry.

The left-hand figure shows that each firm makes q_1^* at the lowest point on its average cost curve LAC_1 . This is also the lowest point on its $SATC$ curve and hence also lies on its SMC curve, though the initial position of these two curves is not shown in Figure 8.8. N_1 firms in the industry each make output q_1^* . Total output is $q_1^* = N_1 q_1^*$.

A rise in input prices raises costs for all firms. LAC_2 is the new long-run average cost curve for a firm. In the short run, a firm has some fixed factors. $SATC_2$ and $SAVC_2$ are average total and average variable costs given these fixed factors. Short-run marginal cost SMC_2 goes through the lowest point of both these curves. The part of SMC_2 above $SAVC_2$ is the firm's short-run supply curve. In the short run, the number of firms remains fixed.

Horizontally adding these short-run supply curves for N_1 firms, we get the new industry short-run supply curve $SRSS_2$. The new short-run equilibrium is at P_2 , where $SRSS_2$ crosses the demand curve. Each firm has $P_2 = SMC_2$ and supplies q_2 . Together, the N_1 firms supply Q_2 . Firms cover variable costs, but not fixed costs, at the price P_2 . They are losing money.

As time elapses, fixed factors are varied, and firms leave the industry. Long-run equilibrium is at price P_2^* since the new long-run industry supply curve $LRSS_2$ is horizontal at P_2^* , which just covers minimum long-run average costs. Each firm supplies q_2^* . The number of firms N_2 is such that $Q_2^* = N_2 q_2^*$.

Figure 8.8 makes two points about the change in long-run equilibrium. First, the rise in average costs is eventually passed on to the consumer in higher prices. In long-run equilibrium the marginal firm (here all firms, since they are identical) breaks even, so there is no incentive for further entry or exit. Hence, price rises to cover the increase in minimum average costs.

Second, since higher prices reduce the total quantity demanded, industry output must fall.

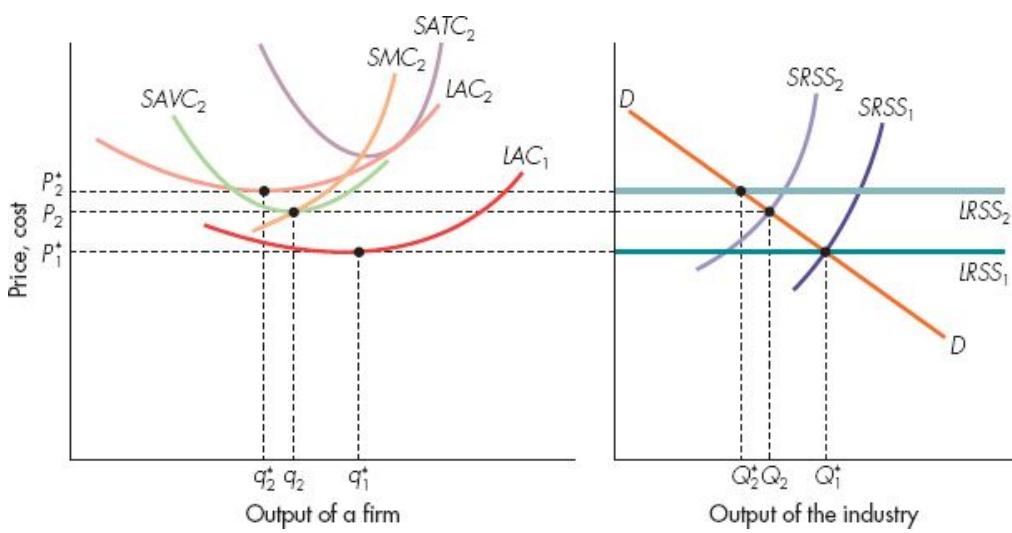


Figure 8.8 A cost increase in a competitive industry

The industry begins in long-run equilibrium producing Q_1^* at a price P_1^* . Each identical firm produces q_1^* at the lowest point on LAC_1 . The long-run supply curve $LRSS_1$ is horizontal at P_1^* . When costs increase, firms have fixed factors and the number of firms is given in the short run. Each firm produces q_2 where the short-run equilibrium price P_2 equals SMC_2 . Together, these firms produce Q_2 . Since firms are losing money, in the long run some firms leave the industry. The new long-run supply curve $LRSS_2$ for the industry is horizontal at P_2^* , the minimum point on each firm's new long-run average cost curve LAC_2 . Each firm produces q_2^* . Industry output is Q_2^* .

A shift in the market demand curve

Figure 8.9 shows the effect of a shift up in the market demand curve from DD to $D'D'$. We show the effects at the industry level. Try to draw your own diagram showing what is happening for the individual firm, as we did in Figure 8.8.

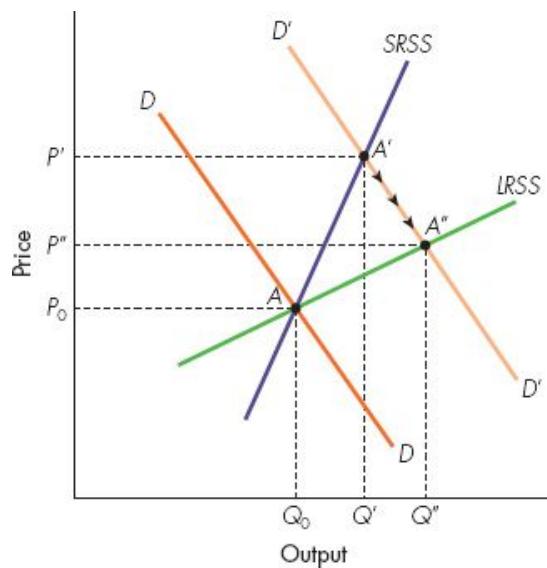


Figure 8.9 A shift in demand in a competitive industry

The industry begins in long-run equilibrium at A. When the demand curve shifts from DD to D'D' the new short-run equilibrium occurs at A'. As fixed factors are gradually adjusted and new firms enter the industry, equilibrium gradually moves from A' towards A'', the new long-run equilibrium.

The industry starts in long-run equilibrium at point A. Overnight, each firm has fixed factors and the number of firms is fixed. Horizontally adding their short-run supply curves, we get the industry supply curve SRSS. The new short-run equilibrium is at A'. When demand first rises, it takes a big price rise to induce individual firms to move up their steep short-run supply curves with given fixed factors.

In the long run, firms can adjust all factors and their long-run supply curves are flatter. Moreover, supernormal profits attract extra firms into the industry. Figure 8.9 assumes that the long-run industry supply curve is rising. Either it takes higher prices to attract higher-cost firms into the industry, or the collective expansion bids up some input prices, or both. The new long-run equilibrium is at A''. Relative to short-run equilibrium at A', there is another rise in total output. However, a better choice of inputs and the entry of new firms raise supply and reduce the market-clearing price.

8.5

Pure monopoly: the opposite limiting case

A perfectly competitive firm is too small to worry about any effect of its output decision on industry supply and hence price. It can sell as much as it wants at the market price. We next discuss the opposite limiting case of market structure, the case of pure monopoly.

A monopolist is the sole supplier in the market. The firm and the industry coincide.³ The sole national supplier may not be a monopolist if the good or service is internationally traded. The Royal Mail is the sole supplier of UK stamps and a **monopolist** in them. Airbus is the only large plane maker in Europe but is not a

monopolist since it faces cut-throat international competition from Boeing. Sole suppliers may also face invisible competition from potential entrants. If so, they are not monopolists.

A **monopolist** is the sole supplier and potential supplier of the industry's product.

First, we study the decisions of a private profit-maximizing monopolist who has no fear of entry or foreign competition. Some monopolies are state-owned and not necessarily run for private profit. However, in the past two decades many countries have been 'privatizing' these state-run monopolies. The analysis in the rest of this chapter is relevant both to existing private monopolies and to how state-run monopolies might behave if restored to private ownership.

8.6 Profit-maximizing output for a monopolist

To maximize profits, any firm chooses the output at which marginal revenue MR equals marginal cost (SMC in the short run and LMC in the long run). It then checks it is covering average costs ($SAVC$ in the short run and LAC in the long run).

The special feature of a competitive firm is that MR equals price. Selling an extra unit of output does not bid down the price and reduce the revenue earned on previous units. The price at which the extra unit is sold is the change in total revenue.

In contrast, a monopolist's demand curve *is* the industry demand curve, which slopes down. Hence MR is less than the price at which the extra output is sold. The monopolist knows that extra output reduces revenue from *existing* units. To sell more, the price on all units must be cut.

In Chapter 6, and in particular Maths 6.1, we explained how, for a downward-sloping demand curve, price, marginal revenue and total revenue are related. Figure 8.10 reminds you of those relationships. The more inelastic the demand curve, the more an extra unit of output bids down the price, reducing revenue from existing units. At any output, MR is further below the demand curve the more inelastic is demand. Also, the larger the existing output, the larger the revenue loss from existing units when the price is reduced to sell another unit. For a given demand curve, MR falls increasingly below price the higher the output from which we begin.

Beyond a certain output (4 in Figure 8.10), the revenue loss on existing output exceeds the revenue gain from the extra unit itself. Marginal revenue is negative. Further expansion reduces total revenue.

On the cost side, with only one producer, the cost curves for a single firm in Chapter 7 carry over directly. The monopolist has the usual cost curves, average and marginal, short run and long run. For simplicity, we discuss only the long-run curves.

Profit-maximizing output

Setting MR equal to MC leads to the profit-maximizing level of positive output. Then the monopolist must check whether, at this output, the price (average revenue) covers average variable costs in the short run and average total costs in the long run. If not, the monopolist should shut down in the short run and leave the industry in the long run. Table 8.2 summarizes the criteria by which a monopolist decides how much to produce.

Table 8.2 Profit-maximizing monopoly

	Marginal condition			Average condition			
				Short run		Long run	
	$MR > MC$	$MR = MC$	$MR < MC$	$P > SAVC$	$P < SAVC$	$P > LAC$	$P < LAC$
Output decision	Raise	Optimal	Lower	Produce	Shut down	Stay	Exit

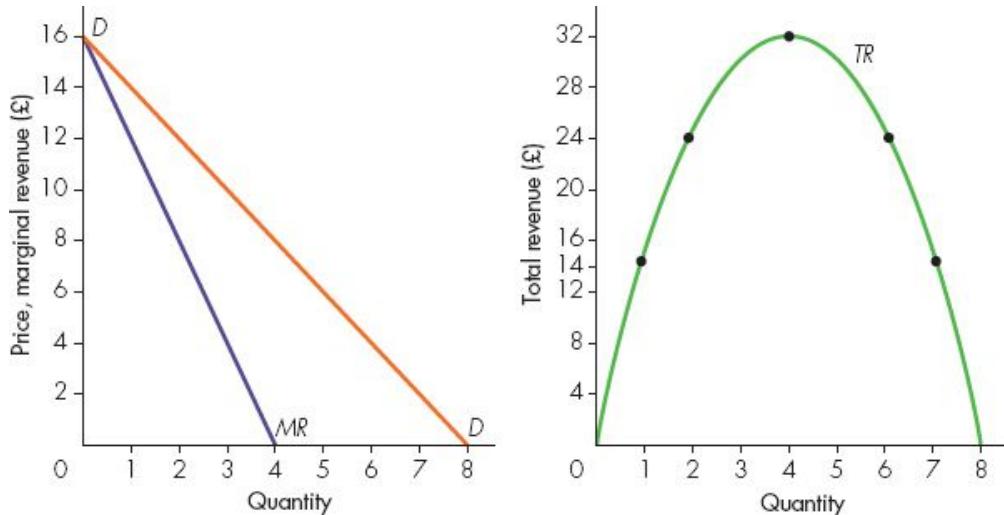


Figure 8.10 Demand, total revenue and marginal revenue

Total revenue (TR) equals price times quantity. From the demand curve DD we can plot the TR curve at each quantity. Maximum TR occurs at £32, when 4 units are sold for £8 each. Marginal revenue (MR) shows how TR changes when quantity is increased a small amount. MR lies below the demand curve DD. From the price of the extra unit we must subtract the loss in revenue from existing units as the price is bid down. This effect is larger the higher is existing output and the more inelastic is the demand curve. The MR curve lies further below DD the larger is output and the more inelastic the demand curve. Beyond an output of 4 units, MR is negative and further expansion reduces total revenue.

Figure 8.11 shows the average cost curve AC with its usual U-shape. The marginal cost curve MC goes through the lowest point on the AC curve. Marginal revenue MR lies below the downward-sloping demand curve DD . Setting $MR = MC$, the monopolist chooses the output Q_1 . To find the price for which Q_1 is sold, we look at the demand curve DD . The monopolist sells output Q_1 at a price P_1 . Profit per unit is $(P_1 - AC_1)$ and total profit is the shaded area $(P_1 - AC_1)Q_1$.

Even in the long run, the monopolist makes *supernormal profits*, sometimes called *monopoly profits*. Unlike the case in competitive industry, supernormal profits of a monopolist are not eliminated by the entry of more firms and a fall in price. A

monopoly has no fear of possible entry. By ruling out entry, we remove the mechanism by which supernormal profits disappear in the long run.

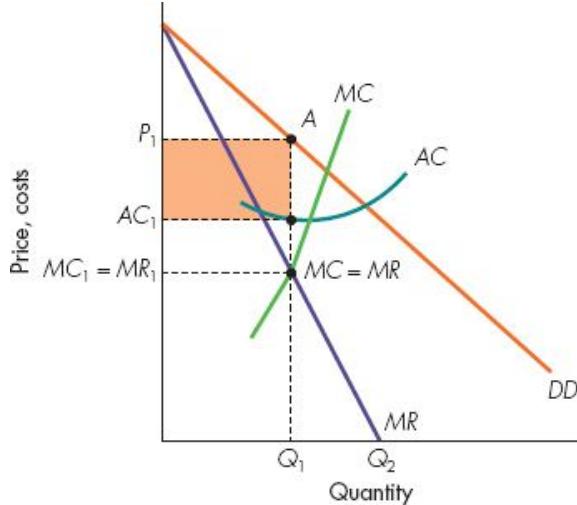


Figure 8.11 The monopoly equilibrium: $MC = MR$

Applying the usual marginal condition, a profit-maximizing monopolist produces the output level Q_1 at which marginal cost MC equals marginal revenue MR . Then it must check that price covers average cost. In this figure, Q_1 can be sold at a price P_1 in excess of average costs AC_1 . Monopoly profits are the shaded area $(P_1 - AC_1) \times Q_1$.

Price setting

Whereas a competitive firm is a *price-taker*, a monopolist sets prices and is a *price-setter*. Having decided to produce Q_1 in Figure 8.11, the monopolist quotes a price P_1 knowing that customers will then demand the output Q_1 .

Elasticity and marginal revenue

When the elasticity of demand is between 0 and 1, demand is inelastic and a rise in output reduces total revenue. Marginal revenue is negative. In percentage terms, the fall in price exceeds the rise in quantity. All outputs to the right of Q_2 in Figure 8.11 have negative MR . The demand curve is inelastic at quantities above Q_2 . At quantities below Q_2 , the demand curve is elastic. Higher output leads to higher revenue. Marginal revenue is positive.

The monopolist sets $MC = MR$. Since MC must be positive, so must MR . The chosen output must lie to the left of Q_2 . *A monopolist never produces on the inelastic part of the demand curve.*

Price, marginal cost and monopoly power

The excess of price over marginal cost is a measure of **monopoly power**.

At any output, price exceeds the monopolist's marginal revenue since the demand curve slopes down. Hence in setting $MR = MC$, the monopolist sets a price that exceeds marginal cost. In contrast, a competitive firm always equates price and marginal cost, since its price is also its marginal revenue. A competitive firm cannot raise price above marginal cost and has no **monopoly power**.⁴ The less elastic the demand curve of a monopolist, the more marginal revenue is below price, the greater is the excess of price over marginal cost, and the more monopoly power it has.

Comparative statics for a monopolist

Figure 8.11 may also be used to analyse changes in costs or demand. Suppose a rise in costs shifts the MC and AC curves upwards. The higher MC curve must cross the MR curve at a lower output. If the monopolist can sell this output at a price that covers average costs, the effect of the cost increase must be to reduce output. Since the demand curve slopes down, lower output means a higher equilibrium price.

Similarly, for the original cost curves shown in Figure 8.11, suppose there is an outward shift in demand and marginal revenue curves. MR must now cross MC at a higher output. Thus a rise in demand leads the monopolist to increase output.

MATHS 8.1

PROFIT MAXIMIZATION AND MONOPOLY POWER

Profit maximization and monopoly power

Consider a monopoly that faces the following linear inverse demand function for its product: $P = a - bQ$, where a and b are positive constants. In order to produce, the monopoly faces the following total cost function: $TC(Q) = cQ + dQ^2$, where c and d are positive constants. The profit function of the monopoly is given by total revenues minus total costs. Total revenues are $TR(Q) = P \times Q$. In our case, the price P is given by the inverse demand function, therefore the total revenues are: $TR(Q) = (a - bQ)Q = -bQ^2 + aQ$.

The profit function of the monopoly is then:

$$\pi(Q) = -bQ^2 + aQ - cQ - dQ^2$$

Equation (1) tells us how the profit of the monopolist (π) depends on the quantity produced (π). The monopolist should choose the quantity to produce in order to maximize the profit function in (1). We know that, at the quantity that maximizes profits, it must be true that $MR = MC$. In our case, we have that:

$$MR(Q) \equiv \frac{dTR(Q)}{dQ} = -2bQ + a \text{ and } MC(Q) \equiv \frac{dTC(Q)}{dQ} = c + 2dQ$$

Therefore the quantity that maximizes profits must solve the equation $MR = MC$:

$$-2bQ + a = c + 2dQ$$

Solving equation (2) for Q gives us:

$$Q^* = \frac{a-c}{2(b+d)}$$

This is the quantity that the monopolist chooses to produce. Notice that we need to assume $a > c$ in order to have a positive quantity produced. Once we know the quantity chosen by the monopolist, we can find the market price. Substitute equation (3) into the inverse demand function:

$$P^* = a - b \left(\frac{a-c}{2(b+d)} \right)$$

A bit of algebra and equation (4) can be written as:

$$P^* = \frac{b(a+c) + 2ad}{2(b+d)}$$

Notice that we have assumed that the monopolist chooses the quantity that maximizes the profit and then we found the price charged using the inverse demand. Obviously, we could have done the reverse. We could have found the price that maximizes the profits (since the monopolist is a price-setter) and then applied that information to the demand function to find the quantity the monopolist should sell. The result would be exactly the same.

Here is a numerical example. Suppose that the demand is $P = 100 - 2Q$, while the total cost function is $TC = 10Q + Q^2$. Compared to our previous analysis, we now have: $a = 100$, $b = 2$, $c = 10$ and $d = 1$. Using those data in (3), we have that the profit-maximizing quantity chosen by the monopolist is $Q^* = 15$, while the price charged by the monopolist is $P^* = 70$. The marginal cost when $Q^* = 15$ is $MC = 40$. As we can see, the price charged by the monopolist is higher than the marginal cost.

Monopoly power measures the ability of the monopolist to set a price higher than marginal cost. The monopoly power crucially depends on the elasticity of demand. To see this more formally, we first relate the marginal revenue to the elasticity of demand. Consider a generic inverse demand function $P(Q)$. The total revenue function is $TR = P(Q) \times Q$. By taking the derivative with respect to Q of TR , we get that the marginal revenue can be written as:

$$MR(Q) = \frac{dP}{dQ}Q + P$$

where dP/dQ is the derivative of the inverse demand function with respect to Q .

Rearrange equation (5) in the following way:

$$MR(Q) = P \left[\frac{\frac{dP}{dQ}Q}{P} + 1 \right]$$

The term $(dP/dQ)(Q/P)$ looks familiar. From Chapter 4 we know that the point elasticity of demand is given by $PED = (dQ/dP)(P/Q)$. Therefore $1/PED = (dP/dQ)(Q/P)$ is the inverse of the elasticity of demand. Using this fact in equation (6), we have:

$$MR(Q) = P \left[1 + \frac{1}{PED} \right]$$

Since the elasticity of demand is a negative number, we can use its absolute value in (7) and we have:

$$MR(Q) = P \left[1 - \frac{1}{|PED|} \right]$$

Equation (8) gives us the relationship between the marginal revenue and the elasticity of demand.

The profit-maximization condition of the monopolist, $MR = MC$, can be written as:

$$P \left[1 - \frac{1}{|PED|} \right] = MC(Q)$$

If demand is inelastic (meaning $|PED|$ between 0 and 1), the term in brackets is negative. There is no way for the left-hand side to be equal to the right-hand side since the marginal cost of production is normally positive. This explains why a monopolist never produces on the inelastic part of the demand curve. Moreover, from equation (9) we can say that the higher is the elasticity of demand (meaning $|PED|$ is a very large positive number), the closer to 1 is the term in brackets. In this case, the price charged by the monopolist will be very close to the marginal cost. Another way to say this is, the more elastic is the demand faced by the monopolist, the lower is its monopoly power.

8.7

Output and price under monopoly and competition

We now compare a perfectly competitive industry with a monopoly. For this comparison to be of interest, the two industries must face the same demand and cost conditions. How would the *same* industry change if it were organized first as a competitive industry then as a monopoly?

Chapter 9 explains why some industries are competitive but others are monopolies. If this theory is right, can the same industry be both competitive and a monopoly? The answer is, only in some special cases.

Comparing a competitive industry and a multi-plant monopolist

Consider a competitive industry in which all firms and potential entrants have the same cost curves. The horizontal *LRSS* curve for this competitive industry is shown in Figure 8.12. Facing the demand curve *DD*, the industry is in long-run equilibrium at *A* at a price P_1 and total output Q_1 . The industry *LRSS* curve is horizontal at P_1 , the lowest point on the *LAC* curve of each firm. Any other price leads eventually to infinite entry to or exit from the industry. *LRSS* is the industry's long-run marginal cost curve LMC_1 of expanding output by enticing new firms into the industry.

Each firm produces at the lowest point on its *LAC* curve, breaking even. The marginal cost curves pass through the point of minimum average costs. Hence, each firm is also on its *SMC* and *LMC* curves. Horizontally adding the *SMC* curves of each firm, we get *SRSS*, the short-run industry supply curve. This is the industry's short-run marginal cost curve SMC_1 of expanding output from existing firms with temporarily fixed factors. Since *SRSS* crosses the demand curve at P_1 , the industry is in both short-run and long-run equilibrium.

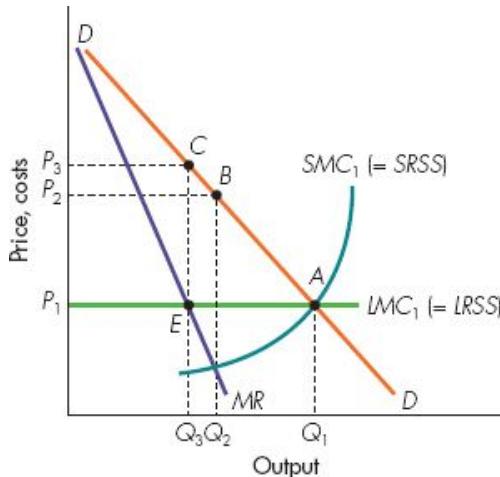


Figure 8.12 A monopolist produces a lower output at a higher price

Long-run equilibrium in a competitive industry occurs at *A*. Total output is Q_1 and the price P_1 . A monopolist sets *MR* equal to SMC_1 , restricting output to Q_2 and increasing price to P_2 . In the long run the monopolist sets *MR* equal to LMC_1 , reducing output to Q_3 and increasing the price again to P_3 . There are no entrants to compete away supernormal profits $P_3 - P_1$ by increasing the industry output.

Beginning from this position, the competitive industry becomes a monopoly. The monopolist takes over each plant (firm) but makes central pricing and output decisions. Overnight, the monopolist still has the same number of factories (ex-firms) as in the competitive industry. Since the firm and the industry now coincide, SMC_1 remains the short-run marginal cost curve for the monopolist, taking all plants together.⁵ However, the monopolist knows that higher total output bids down the price.

In the short run, the monopolist equates SMC_1 and *MR*, reaching equilibrium at *B*. Output is Q_2 and the price is P_2 . Relative to competitive equilibrium at *A*, the monopolist raises price and reduces quantity.

In the long run, the monopolist can enter (set up new factories) or exit (close down existing factories). Whether making short-run profits or losses at B (we need to draw the $SATC$ curve to see which), a monopolist will now exit the industry or close down some factories in the long run.

The monopolist cuts back output to force up the price. In the long run, it makes sense to operate each factory at the lowest point on its LAC curve. To reduce total output some factories are closed. In the long run, the monopolist sets $LMC_1 = MR$ and reaches equilibrium at C . Price has risen yet further to P_3 and output has fallen to Q_3 . Long-run profits are given by the area P_3CEP_1 , since P_1 remains the long-run average cost when all plants are at the lowest point on their LAC curve.

Because MR is less than price, a monopolist produces less than a competitive industry and charges a higher price. However, in this example it is a legal prohibition on entry by competitors that allows the monopolist to succeed in the long run. Otherwise, with identical cost curves, other firms would set up in competition, expand industry output and compete away these supernormal profits. Absence of entry is intrinsic to the model of monopoly.

The social cost of monopoly

We know that a monopoly tends to produce less and at a higher price than a perfectly competitive industry. This may be seen as a bad thing, since we should prefer more production at a lower price than less production at a high price (unless of course you are the monopolist). Another way to see why a monopoly may be bad compared to perfect competition is to look at the total surplus in the market, defined as the sum of the consumers' and producers' surplus. From Chapter 3 we know that the total surplus, or social surplus, in the market is a measure of the gain that the participants get from participating in the market.

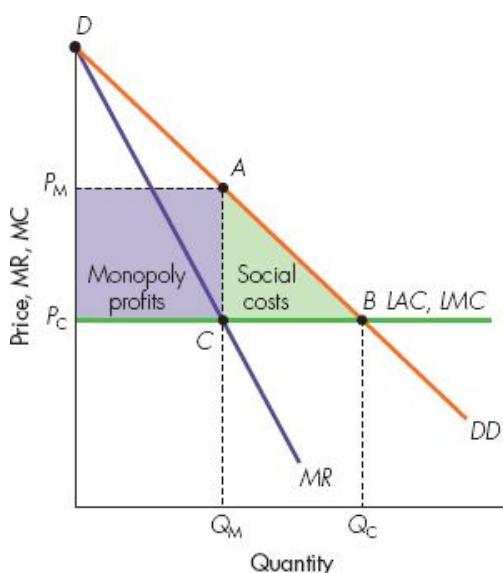


Figure 8.13 The social cost of monopoly

The industry has horizontal long-run average and marginal costs. A perfectly competitive industry produces at B , but a monopolist sets $MR = MC$ to produce only Q_M at a price P_M . The monopolist earns excess profits $P_M P_C CA$, but there is a social cost or deadweight burden equal to the triangle ACB . Between Q_M and Q_C social marginal benefit exceeds social marginal cost and society would gain by expanding output to Q_C . The triangle ACB shows how much society would gain by this expansion.

Consider for simplicity an industry with a constant long-run marginal cost equal to long-run average cost. An example of a cost function with such a property is $TC = cQ$, where c is a positive constant. In Figure 8.13, under perfect competition LMC is both the industry's long-run marginal cost curve and its supply curve. With constant returns to scale, LMC is also the long-run average cost curve of the industry. Given the demand curve DD , competitive equilibrium is at B . The competitive industry produces an output Q_C at a price P_C .

Now the industry becomes a monopolist, producing output Q_M at a price P_M , thus equating marginal cost and marginal revenue. Under a perfectly competitive industry, the producer's surplus is zero and the total surplus coincides with the consumers' surplus; this is given by area $P_C BD$ in Figure 8.13.

Under a monopoly industry, the producer's surplus is given by area $P_M P_C AC$, which represents the monopolist's profits. On the other hand, the consumers' surplus is now given by area $P_M AD$. The total surplus under a monopoly is therefore area $P_C CAD$.

By comparing the case of a competitive industry with the monopoly, we can see that under a monopoly the total surplus is lower than under a perfectly competitive industry.

At an output below the efficient level, the **deadweight loss** shows the loss of social surplus.

In particular, the difference between the two cases is represented by the triangle ACB in Figure 8.13. That triangle represents the **deadweight loss** in social surplus caused by the fact that the monopolist restricts output to Q_M . Notice that this deadweight loss is just a waste in social surplus. By increasing the level of competition in the industry, it is possible to increase the quantity produced and to reduce that loss. When the industry becomes perfectly competitive, in the sense that output will be sold at a price equal to the long-run marginal cost, the social surplus is maximized, meaning that it cannot be increased further.

For the economy, the social cost of monopoly is found by adding together the deadweight loss triangles such as ACB for all industries.

The **social cost of monopoly** is the failure to maximize social surplus.

Is the social cost of monopoly power large? Economists who believe in market forces tend to think it is small. Professor George Stigler, a Nobel Prize winner, once quipped, 'Economists might serve a more useful purpose if they fought fires or termites instead of monopoly.' Other economists believe the **social cost of monopoly** is much larger.

Why such a disagreement? First, the area of the deadweight loss triangle in Figure 8.13 depends on the elasticity of the demand curve. In calculating the size of deadweight loss triangles under monopoly, different economists use different estimates of the demand elasticity.

Second, the welfare cost of monopoly is not just the deadweight loss. Since monopoly may yield high profits to the firm, firms spend a lot trying to acquire and secure monopoly positions. Firms may devote large quantities of resources trying to influence the government in ways that enhance or preserve their monopoly power. They may also deliberately maintain extra production capacity to create a credible threat to flood the market if an entrant comes in. Socially, resources devoted to lobbying the government or maintaining overcapacity are largely wasted. This kind of behaviour by firms to acquire and protect monopoly profits is called *rent-seeking* behaviour and it may be socially wasteful.

CASE 8.1

MONOPOLY POWER AND COMPETITION POLICY

We have seen that a monopoly creates a social loss compared to a perfectly competitive market. If it is possible to increase the level of competition in a monopolized market, then society is better off since social surplus increases. Competition policy (also known as antitrust policy) deals with markets where competition can arise; however, given the behaviour of some firms in those markets, competition is restricted. There are markets in which increasing the level of competition is not feasible, so competition policy does not apply. This is the case of a natural monopoly, which will be discussed at the end of this chapter.

Broadly speaking, competition policy can be divided into policies to deal with monopoly power that already exists, and policies to deal with mergers that may increase monopoly power. While mergers will be discussed in the next chapter, here we discuss policies to address existing monopoly power. Since the UK belongs to the European Union, EU competition law takes precedence where it is relevant, essentially in the case of larger businesses with significant European or global activities.

The original Common Market was created by the 1956 Treaty of Rome. The modern and enlarged EU is largely underpinned by the 1999 Treaty of Amsterdam. Article 81 of this treaty prohibits anti-competitive agreements (called cartels) that have an appreciable effect on trade between EU member states and which prevent or distort competition within the EU.⁶ Article 82 prohibits the abuse of any existing dominant position. A firm has a dominant position in a given market if it has a large market share in that market. For example, Microsoft has a dominant position in the market for operating systems (OS) for PCs, with a market share of around 90 per cent. Article 82 prohibits the abuse of a dominant position not the dominant position itself. A firm can become a dominant firm simply because it is more productive than the others and this is fine for competition policy. What is not fine is a firm that uses its dominant position to restrict competition in the market.

Responsibility for enforcement of these articles lies with the European Commission.

Although global businesses are increasingly subject to transnational competition law, many businesses still operate primarily within one country; national decisions are then appropriate. Within the UK, these are governed by the Competition Act 1998 and the Enterprise Act 2002. The latter made it a criminal offence, punishable by a jail sentence, to engage in a dishonest cartel.

Two key institutions addressing UK competition policy are the Office of Fair Trading (OFT) and the Competition Commission. In particular, the OFT has the power to refer cases in which existing monopoly power may be leading to a ‘substantial lessening of competition’ to the Competition Commission for detailed investigation.

Prior to the Enterprise Act 2002, the Competition Commission was asked instead to evaluate whether or not a monopoly was acting ‘in the public interest’, without any presumption that monopoly was bad, and many previous judgements of the Commission concluded that companies were acting in the public interest, for example because they had an excellent record of innovation, *despite* having a monopoly position.

The change in 2002 therefore emphasized competition more strongly and made the Competition Commission more accountable by defining its objectives more clearly. This also brought UK law more clearly into line with EU competition law, by placing measures of competition at the centre of the evaluation of competition policy.

8.8 A monopoly has no supply curve

A competitive firm sets price equal to marginal cost if it supplies at all. If we know its marginal cost curve, we know how much it supplies at each price. Aggregating across firms, we also know how much the industry supplies at each price. We can draw the supply curve without knowing anything about the market demand curve. We then analyse how supply and demand interact to determine equilibrium price and quantity.

A monopolist’s output affects marginal cost and marginal revenue simultaneously. Figure 8.14 shows a given *LMC* curve. How much will the monopolist produce at the price P_1 ? It all depends on demand and marginal revenue. When demand is *DD*, marginal revenue is *MR* and the monopolist produces Q_1 at a price P_1 . If demand is *D'D'*, marginal revenue is *MR'*, and the monopolist produces Q_2 but still charges P_1 .

A monopolist does not have a supply curve independent of demand conditions. Rather, a monopolist simultaneously examines demand (hence marginal revenue) and cost (hence marginal cost) to decide how much to produce and what to charge.

Discriminating monopoly

A **discriminating monopoly** charges different prices to different people.

Thus far we have assumed that all consumers are charged the same price. Unlike a competitive industry, where competition prevents any firm charging more than its competitors, a monopolist may be able to charge different prices to different customers. When a producer charges different customers different prices for the same product and this difference is not justified by differences in costs, we say that the producer *price discriminates*.

Consider an airline monopolizing flights between London and Rome. It has business customers whose demand curve is not very elastic. They have to fly. Their demand and marginal revenue curves are very steep. The airline also carries tourists whose demand curve is much more elastic. If flights to Rome get too expensive, tourists may visit Athens instead. Tourists have much flatter demand and marginal revenue curves.

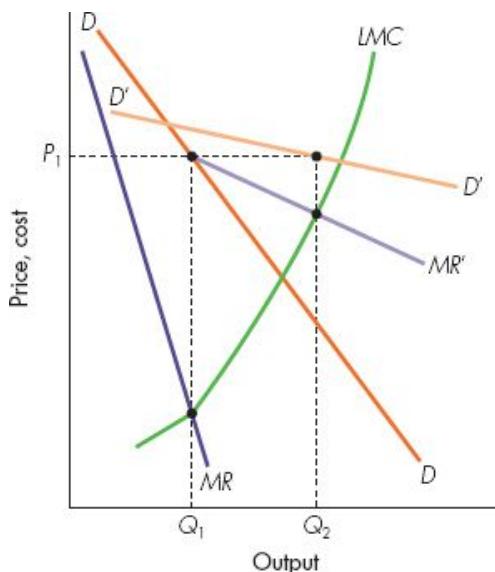


Figure 8.14Absence of a supply curve under monopoly

Given the demand curve DD and the corresponding marginal revenue curve MR, the monopolist produces Q_1 at a price P_1 . However, lacking DD and MR, the monopolist produces Q_2 at a price P_1 . Knowing the price, we cannot uniquely infer the quantity supplied unless we also know demand and marginal revenue. Because the monopolist knows that output affects both marginal cost and marginal revenue, the two must be considered simultaneously.

The less elastic is the demand curve, the more the marginal revenue curve lies below the demand curve. To sell another output unit requires a bigger price cut that hits existing revenue. Since demand elasticity determines the gap between marginal revenue and price, charging the same price to purchasers with different demand elasticities means that the marginal revenue from the last business traveller is less than the marginal revenue from the last tourist.

Whatever the total number of passengers (and total cost of carrying them), the airline then has the wrong *mix* between tourists and business travellers. Since the marginal revenue from the last tourist exceeds the marginal revenue from the last business traveller, the airline gains revenue with no extra cost by carrying one more tourist and one fewer business traveller. It pays to keep changing the mix until the marginal revenue from the two types is equal.

To do this, the airline must charge the two groups *different* prices. Since tourist demand is elastic, the airline charges tourists a low fare to raise tourist revenue. Since business demand is less elastic than tourist demand, the airline charges business travellers a high fare to increase business revenue. This kind of price discrimination is called *third-degree price discrimination*. In this case, the monopoly can divide its customers into different groups according to some characteristics of those groups that affect their demand (for example, business travellers versus tourist travellers, old customers versus young customers, male customers versus female customers, and so on).

Profit-maximizing output satisfies two separate conditions. First, business travellers with inelastic demand pay sufficiently more than tourists with elastic demand so that the marginal revenue from the two types is equal. There is thus no incentive to rearrange the mix by altering the price differential between the two groups. Second, the level of prices and the total number of passengers is determined to equate marginal cost to each of these marginal revenues. The airline operates at the most profitable scale as well as with the most profitable mix. If a monopolist can price discriminate, it is profitable to do so. Charging different customers different prices for the same product is more profitable than charging all customers the same price.

The case of third-degree price discrimination is illustrated in Figure 8.15. Suppose that the monopolist can ‘segment’ the market into two groups of customers (such as business travellers versus tourists). The demand of customers in group 1, DD_1 , is more elastic than the demand of customers in group 2, DD_2 . Customers in group 2 have a higher willingness to pay for the good than customers in group 1. For simplicity, the marginal cost of production is constant and it is denoted by horizontal line MC . Profit maximization implies that the price charged to each group of customers satisfies marginal revenue equal to marginal cost; therefore, $MR_1 = MC = MR_2$. The monopolist sets the prices it charges the two groups in such a way that the marginal revenues from both groups are equal. Since group 1 has higher demand elasticity than group 2, for the marginal revenues to be equal across groups, P_2 must be higher than P_1 . The monopolist charges a higher price to the customers with low elasticity of demand. Profit-maximizing quantities are Q_1^* and Q_2^* .

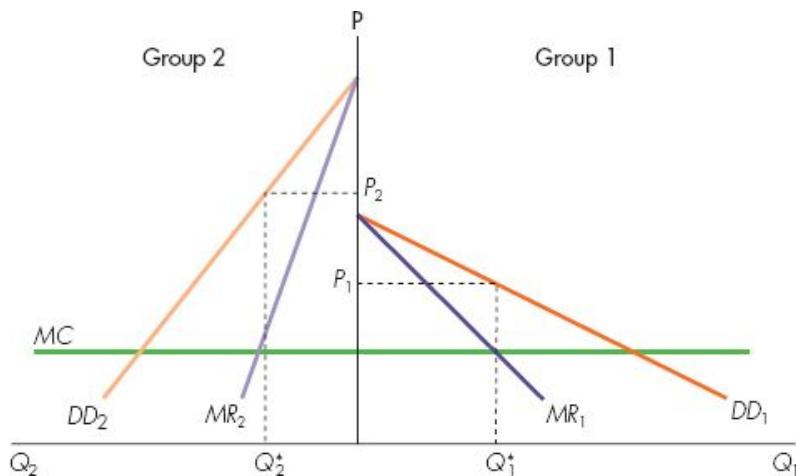


Figure 8.15 Third-degree price discrimination

If the monopolist can divide its customers into two different groups it will charge the customers with a low elasticity of demand (Group 2) a higher price than the customers with high elasticity of demand (Group 1). The monopolist is price discriminating. Profit maximization implies that the monopolist will set the price for each group such that marginal revenues of each group are the same and equal to the marginal cost.

There are many examples of third-degree price discrimination in the real world. Rail companies charge rush-hour commuters a higher fare than midday shoppers whose demand for trips to the city is much more elastic. Night clubs may charge different entry fees to men and women. Ticket prices for cinemas and public transport are normally different for adults and children.

Most examples of price discrimination refer to services consumed on the spot rather than to goods that can be resold. Price discrimination in a standardized commodity is unlikely to work. Those buying at the low price resell to those paying the high price, undercutting price discrimination. Effective price discrimination requires that resale between the sub-markets is not feasible.

Price discrimination illustrates again the absence of a supply curve under monopoly. Figure 8.16 shows *perfect price discrimination*, also known as *first-degree price discrimination*. Each customer pays a different price for the same product. In particular, each customer pays according to his willingness to pay for the good, which is the maximum price at which he is willing to buy.

If a monopolist charges every customer the same price, profit-maximizing output is Q_1 , where MR equals MC and the price is P_1 .

If the monopolist can perfectly price discriminate, the very first unit of output can be sold at a price E . Having sold the first unit to the highest bidder most desperate for the good, the next unit can be sold to the next-highest bidder, and so on. Moving down the demand curve DD , we can read off the price for which each extra unit is sold. In reducing the price to sell extra output, the monopolist no longer reduces revenue from previously sold units. *Hence the demand curve is the marginal revenue curve under perfect price discrimination.* The marginal revenue of the last unit is simply the price for which it is sold.

Treating DD as the marginal revenue curve, a perfectly price-discriminating monopolist produces at point C , where marginal revenue and marginal cost are equal. Two points follow immediately. First, if price discrimination is possible, it is profitable to use it. Moving from the uniform pricing point A to the price discriminating point C , the monopolist adds the area ABC to profits. This is the excess of extra revenue over extra cost when output is increased.

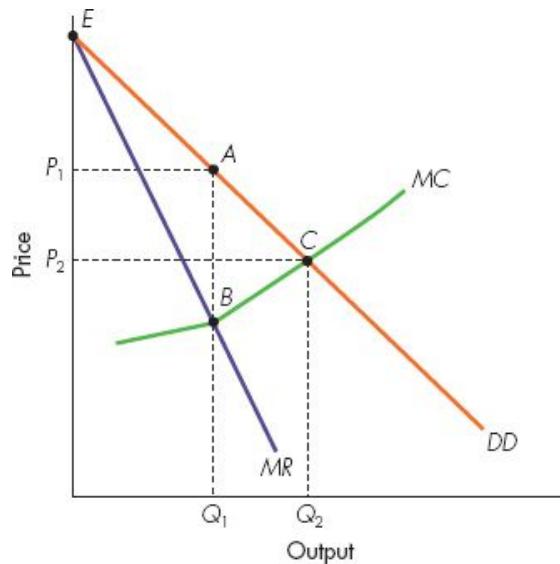


Figure 8.16Absence of a supply curve under monopoly

Charging all customers the same price the monopolist will produce at B where $MC = MR$. If each output unit can be sold for a different price the revenue from existing units is not reduced by cutting the price to sell another unit. The demand curve DD is the marginal revenue curve and the perfectly discriminating monopolist will produce at C . Output is higher and profits are higher. By price discrimination the monopolist gains an extra revenue EP_1A from selling Q_1 but also increases output beyond this level making a marginal profit of ABC in expanding from Q_1 to Q_2 .

The monopolist makes a second gain from price discrimination. Even the output Q_1 now earns more revenue than under uniform pricing. The monopolist also gains the area EP_1A by charging different prices for the first Q_1 units of output rather than the single price P_1 . In practice, one of the main ways management consultants raise the profits of firms that they advise is by devising new ways in which the firm can price discriminate.

Second, whether or not the firm can price discriminate affects its output choice. Uniform pricing leads to an output Q_1 . Perfect price discrimination leads to an output Q_2 . Uniform and discriminatory pricing lead to different outputs because they affect the marginal revenue obtained by a monopolist from a given demand curve.

Perfect price discrimination is normally not feasible in reality. In order to perfectly discriminate, the monopolist should know the willingness to pay of each consumer – information that is impractical to obtain. Nevertheless, perfect discrimination gives an interesting result. If the monopolist can perfectly discriminate, the social loss associated with a monopoly disappears. The social surplus under a monopoly that perfectly

discriminates is the same as under perfect competition. However, the distribution of the social surplus is completely different in the two cases. Under perfect competition the producer surplus is zero, while the social surplus coincides with consumers' surplus. For a perfectly discriminating monopolist, the reverse is true. The perfectly discriminating monopolist extracts the entire consumers' surplus, and so the social surplus coincides with the producer surplus (the profits of the monopolist).

Finally, a monopolist may discriminate on the quantity sold to customers. Even if the firm is not able to determine the willingness to pay of its customers, it knows that the demand of each customer is normally negatively sloped. This means that customers should have higher willingness to pay for the first units of a good. A monopolist may discriminate by allowing the price a customer pays to vary with the quantity purchased. Customers buying different quantities will be charged different unit prices. A typical example is a quantity discount. This case of price discrimination is called *second-degree price discrimination* or quantity discrimination.

8.9 Monopoly and technical change

Section 8.7 compared a monopoly and a perfectly competitive industry. When such a comparison was meaningful, we discovered: (1) a monopoly will restrict output and drive up prices, and (2) a monopoly may make economic profits permanently.

The economist and social scientist, Joseph Schumpeter (1883–1950), argued that this comparison ignores technical advances that reduce costs, allowing price cuts and output expansion. A large monopolist with steady profits may find it easier to *fund* the research and development (R&D) necessary to make cost-saving breakthroughs. More importantly, a monopolist may have more *incentive* to undertake R&D.

In a competitive industry a firm with a technical advantage has only a temporary opportunity to earn high profits to recoup its research expenses. Imitation by existing firms and new entrants competes away profits. In contrast, by shifting all its cost curves downwards, a monopoly can enjoy higher profits for ever. Schumpeter argued that these two forces – more resources for R&D and a higher return on a successful venture – make monopolies more innovative than competitive industries. Taking a dynamic long-run view, not a static picture, monopolists enjoy lower cost curves. As a result, they charge lower prices, thus raising the quantity demanded.

This argument has some substance. Tiny firms often do little R&D. Many of the largest firms have excellent research departments. Even so, the Schumpeter argument may overstate the case.

Modern economies have a *patent* system. Inventors of new processes get a temporary legal monopoly for a fixed period. By temporarily excluding entry and imitation, patent laws raise the incentive to conduct R&D but do not establish a monopoly in the long run. Over the life of the patent, the inventor charges a higher price and makes handsome profits. Eventually, the patent expires and competition from other firms leads to higher output and lower prices. The real price of copiers and microcomputers fell significantly when the original patents of Xerox and IBM expired.

8.10 Natural monopoly

A **natural monopoly**'s average costs keep falling as its output rises. It undercuts all smaller competitors.

Consider a monopolist meeting the entire industry demand from a single plant. This is most plausible when scale economies are big. There are huge costs in setting up a national telephone network, for example. Yet the cost of connecting a marginal subscriber is low once the network has been set up.

Monopolies enjoying huge economies of scale – falling *LAC* curves over the entire range of output and always above the *LMC* curve – are **natural monopolies**. Large-scale economies may explain why there is a sole supplier with no fear of entry by others. Smaller new entrants would be at a prohibitive cost disadvantage.

Figure 8.17 shows an industry with steadily falling long-run average costs as output rises. Only one private firm can survive in such an industry. Any firm that expands output can cut costs and undercut its rivals. Facing a demand curve *DD* and marginal revenue curve *MR*, the resulting monopolist produces Q_M and earns profits $P_M CBE$. The monopolist makes too little and it creates a deadweight burden AEE' .

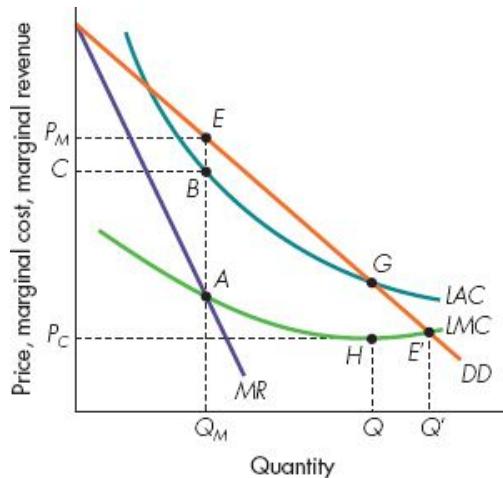


Figure 8.17 Natural monopoly

The efficient point E' equates long-run marginal cost LMC and marginal benefit DD . A private monopolist sets $MR = MC$, produces Q_M and earns profits $P_M CBE$. The deadweight loss under private monopoly is AEE' . If by law the monopolist was forced to charge a fixed price P_C , the monopolist would face a horizontal demand curve $P_C E'$ up to the output Q' . Since P_C would then also be marginal revenue, the monopolist would produce at E' where the marginal revenue and marginal cost coincide. Although efficient, society cannot force the monopolist to produce here in the long run. Since E' lies below LAC the monopolist is making a loss and would rather go out of business.

In the case of a natural monopoly, competition cannot be increased in the market. It is better to have a single firm producing in the market. The reason is that, if you split up the firm to create competition, a lot of small firms each produce at higher average cost and this is a waste of society's resources.

You could order the firm to produce at the efficient point E' . You will get the desired output Q' , but the price P_C is below the firm's average costs at Q' . It makes losses. Since marginal cost always lies below average cost when average cost is falling, forcing a natural monopolist to price at marginal cost is always loss-making. You cannot force a private firm to make losses. It will shut down.

One solution is the use of *regulation* (see also Concept 8.2). In the UK, for example, Ofgem is the regulatory body for the gas and electricity markets. It aims to get close to the efficient allocation E' while letting the monopolist break even after allowing a proper deduction for all economic costs. By making the monopolist produce Q at the price corresponding to average cost at this output, the deadweight burden is cut from AEE' to GHE' .

A **two-part tariff** charges a fixed sum for access to the service and then a price per unit that reflects the marginal cost of production.

An even better solution is to allow the monopolist to charge a **two-part tariff**.

A two-part tariff uses fixed charges to pay for fixed costs, and marginal charges to cover marginal costs. For example, for a landline telephone a customer will be charged a fixed monthly connection fee plus a price per minute of use. The fixed connection fee relates to the fixed costs of the telephone company while the price paid each time a phone call is made relates to the marginal costs.

In Figure 8.17 the monopolist is told to charge P_C for each unit of the good. Consumers demand the socially efficient quantity Q' . Since the monopolist is now a price-taker at the controlled price P_C , it is loss-minimizing for the monopolist to produce Q' , at which both price and marginal revenue equal marginal cost. The regulator then allows the monopolist to levy the minimum fixed charge necessary to ensure that it breaks even after allowing for all relevant economic costs.

A third solution to the natural monopoly problem is to order the monopolist to produce at the efficient point E' at the price P_C , and for the government to provide a subsidy to cover the losses entailed. It is socially desirable to make the efficient output Q' in the cost-minimizing way. If the subsidy solution is adopted, there is pressure for the government to get involved in the entire running of the industry so that operations can be carefully monitored.

Three problems recur with all these solutions to the problem of natural monopoly. First, information is costly for monitors to acquire. It is hard to ensure that the industry strives to keep its cost curves at their lowest possible positions. Unnecessarily high costs can be passed on under average cost pricing (solution 1), can result in a higher fixed charge to ensure break-even under a two-part tariff (solution 2), or can require a larger subsidy

(solution 3). In each case, the regulatory body has the difficult task of trying to make the natural monopoly as efficient as possible.

Regulatory capture implies that the regulator gradually comes to identify with the interests of the firm it regulates, becoming its champion not its watchdog.

The second problem is **regulatory capture**. Regulated companies devote considerable time, effort and money to lobbying the regulator. Of necessity, regulators build up contacts with the regulated. Eventually, the regulator can come to sympathize with the problems of the regulated.

Third, regulators find it hard to make credible commitments regarding their future behaviour. For example, the regulator may encourage the monopolist to invest by promising ‘light’ regulation in the future. Once the investment is made and the cost sunk, the regulator then faces temptations to change the ground rules, toughening requirements. Foreseeing all this, the monopolist does not invest in the first place. There is underinvestment if the regulator faces commitment problems.

During 1945–80, many European governments concluded that the least-bad solution to these problems was nationalization of natural monopoly industries.

CONCEPT 8.2

REGULATION OF NATURAL MONOPOLIES

Natural monopolies were normally state-owned firms. Industries like electricity, gas, water, telecommunications and the railways are all examples of natural monopolies that were nationalized. In those industries, the price charged by the state-owned firms was regulated. This means that those natural monopolies were not allowed to charge the monopoly price for the good they were selling.

There are various ways in which a price can be regulated:

First-best solution: $P = MC$. The first possibility is to force the natural monopoly to charge a price equal to its marginal cost. This possibility represents a first-best solution since it mimics the same result under perfect competition and so no deadweight loss will arise in the market. However, this solution implies that the natural monopoly faces a loss all the time. In this case, the government (the owner of the natural monopoly) must subsidize the natural monopoly.

Second-best solution: $P = AC$. In this case, the natural monopoly is forced to charge a price equal to the average cost (like point *G* in Figure 8.16), and faces zero profits all of the time. The government does not need to subsidize the firm but a deadweight loss is still present in the market. This kind of regulation has been used extensively in real-world cases and is known as *rate of return regulation*.

The main problem with this kind of regulation is that a natural monopoly has no incentive to be efficient. This regulation is a low-power incentive mechanism. Even if the natural monopoly is efficient and can effectively reduce its costs, the

price charged will always reflect the average cost and so the natural monopoly will always get zero profits.

Price cap regulation: the regulator sets the maximum price (price cap) that the firm can set. The firm is free to charge any price equal to or below the maximum price. In contrast to the average cost pricing regulation, a price cap regulation is a high-power incentive mechanism since it provides incentives for the regulated firm to reduce its costs. The regulated firm can charge the maximum price allowed and, by reducing its costs, can increase its profits.

In the UK the price cap was defined according to the formula $RPI - X$. Nominal prices could rise by the same percentage as the retail price index (*RPI*), minus X per cent. X is the annual cut in *real* prices. For example, this kind of regulation was used when British Telecom (BT) was privatized in 1984. Since telecommunications enjoy rapid technical progress, BT could reduce costs year on year in real terms. Initially, the regulator set X at 3 per cent a year, but later raised it to 4.5 per cent and then 6.25 per cent. During its first ten years as a private company, BT cut its real price by 43 per cent.

Summary

- In a **competitive industry** each buyer and seller is a **price-taker**, believing individual actions have no effect on the market price. Competitive supply is most plausible when many firms make a standard product, with free entry and exit, and easy verification by buyers that the products of different firms really are the same.
- For a **competitive firm**, the price is its marginal revenue. Output equates price to marginal cost. The firm's supply curve is its *SMC* curve above *SAVC*. At a lower price, the firm temporarily shuts down. In the long run, the firm's supply curve is its *LMC* curve above its *LAC* curve. At a lower price, the firm eventually exits the industry.
- Adding at each price the quantities supplied by each firm, we obtain the **industry supply curve**. It is flatter in the long run both because each firm can fully adjust all factors and because the number of firms in the industry can vary. In the extreme case where all potential and existing firms have identical costs, the long-run industry supply curve is horizontal at the price corresponding to the lowest point on each firm's *LAC* curve.
- An increase in demand leads to a large price increase but only a small increase in quantity. The existing firms move up their steep *SMC* curves. Price exceeds average costs and the ensuing profits attract new entrants. In the long run, output increases still further but the price falls back. In long-run equilibrium, the **marginal firm** makes only normal profits and there is no further change in the number of firms in the industry.

- An increase in costs for all firms reduces the industry's output and increases the price. In the long run, the marginal firm must break even. A higher price is required to match the increase in its average costs.
- **pure monopoly** is the only seller or potential seller of a good and need not worry about entry, even in the long run. Though rare in practice, this case offers an important benchmark against which to compare less extreme forms of monopoly power.
- **profit-maximizing monopolist has a supply rule** – choose output to set MC equal to MR – but not a supply curve uniquely relating price and output. The relationship between price and MR depends on the demand curve.
- Where a monopoly and a competitive industry can meaningfully be compared, the monopolist produces a smaller output at a higher price. Compared to a perfectly competitive market, a monopoly creates a **deadweight loss**. This is the loss in social surplus caused by the monopolist's restriction of output compared to perfect competition.
- **discriminating monopolist** charges different prices to different customers. To equate the marginal revenue from different groups, groups with high elasticity of demand must pay a lower price. Successful price discrimination requires that customers cannot trade the product among themselves.
- Monopolies may have more internal resources available for research and may have a higher incentive for cost-saving research because the profits from technical advances will not be eroded by entry. Although small firms do not undertake a great deal of expensive research, it appears that the **patent laws** provide adequate incentives for medium- and larger-sized firms. There is no evidence that an industry has to be a monopoly to undertake cost-saving research.

Review questions



EASY

- 1 A competitive industry has free entry and exit. Why does free exit matter? How would the analysis change if it was costly to exit?
- 2 We rarely see a perfectly competitive market because all the assumptions underlying competitive markets rarely hold together in reality. Why do we need to study something that may not exist in the real world?
- 3 Compare perfect competition and monopoly on the basis of:
 - a. the number of buyers and sellers.
 - b. the market supply curve.
 - c. the nature of the good sold in the market.
- 4 **True or False** (a) In a monopoly market, the social welfare is always lower than in a competitive market. (b) Price discrimination is likely to be most effective when the good being sold is a standardized commodity. (c) A firm charges different prices to

customers buying different quantities. This is an example of third-degree price discrimination.

- 5 Common fallacies** Why are these statements wrong? (a) Since competitive firms break even in the long run, there is no incentive to be a competitive firm. (b) By breaking up monopolies, we always get more output at a lower price.

MEDIUM

- 6 The following table reports the data on total costs of a competitive firm. We know that the market price is $P = 44$. Find the marginal cost curve. In a graph, plot the marginal revenue and marginal cost curves and show the amount of output that the firm should produce.

Q	1	2	3	4	5	6	7	8	9
TC	4	16	36	64	100	144	196	256	324

- 7 Draw a diagram showing a competitive industry in short-run equilibrium. Suppose this is the wool industry. The development of artificial fibres reduces the demand for wool. (a) Show what happens in the short run if all sheep farmers have identical costs. (b) What happens in the long run if there are high-cost and low-cost sheep farmers in the industry?
- 8 The table below shows the demand curve facing a monopolist who produces at a constant marginal cost of £6. Calculate the monopolist's marginal revenue curve. What is the equilibrium output? What is the equilibrium price?

Price (£)	8	7	6	5	4	3	2	1	0
Quantity	1	2	3	4	5	6	7	8	9

- 9 The table below shows the demand curve facing a monopolist who produces at a constant marginal cost of £6. Now suppose that, in addition to the constant marginal cost of £6, the monopolist has a fixed cost of £2. How does this affect the monopolist's output, price and profits? Why?

Price (£)	8	7	6	5	4	3	2	1	0
Quantity	1	2	3	4	5	6	7	8	9

- 10 A monopolist faces the following inverse market demand: $P = 50 - 2Q$. Suppose that the total cost faced by the monopolist is $TC = 10Q$. Find the profit-maximizing quantity produced by the monopolist. What about the price charged by the monopolist? Find the deadweight loss in the market. Illustrate your answer in a diagram.
- 11 The following table reports the total cost for a natural monopoly: Find the average cost curve and the marginal cost curve and plot them on a graph. What is the relationship between the two curves?

Q	1	2	3	4	5	6	7	8	9
TC	22	24	26	28	30	32	34	36	38

HARD

- |2 Consider a perfectly competitive firm that has a total cost of producing output given by: $TC = 10Q + 2Q^2$. The market price is $P = 54$. Find the profit-maximizing quantity produced by the firm.

- |3 Suppose that the total output produced in a competitive market is 200 units. Suppose there are n identical firms in the market. Each firm then produces an amount $200/n$.

The total cost of a single firm in the market is If the market $TC = \left(\frac{200}{n}\right)^2$ price is $P = 10$, find the number of firms active in the market.

- |4 A firm's market power can be measured by its ability to raise price above marginal cost. Relative to the level of marginal cost, this measure is $(P - MC)/P$. How do you expect this to be related to the elasticity of demand for the monopolist's output?

-
- 1 Market structure normally refers to the number of firms operating in a market so it is determined by the supply side. The main case in which the demand side determines the market structure is considered in Chapter 10, when the monopsony market is discussed.
 - 2 With this assumption on the demand side, the terms 'industry' and 'market' becomes synonymous and can be used interchangeably. A perfectly competitive industry must operate in a perfectly competitive market. Similarly, a monopoly industry and a monopoly market mean the same thing. We use the term 'industry' throughout the chapter to emphasize the fact that we are focusing on the supply side of a market.
 - 3 Another way to say it is: a monopolist has 100 per cent of market share. A firm's market share is defined as the firm's quantity produced divided by the total quantity produced in the market and then multiplied by 100.
 - 4 A synonym of monopoly power is market power, so those terms can be used interchangeably. Here, we use the expression monopoly power since we are dealing with a monopolist. However, there may be firms that are not monopolists in their market that may still have the ability to set a price above their marginal cost. In those cases, we say that those firms have market power. This will be discussed in detail in Chapter 9.
 - 5 In a competitive industry each firm equates the price to its own marginal cost. Hence firms produce at the same marginal cost. We horizontally add individual SMC curves (at the same price) to get the industry SMC curve. A multi-plant monopolist need not equate MC across all plants but always finds it profitable to do so. If marginal costs in two plants differ, a monopolist can produce the same total output more cheaply by producing an extra unit in the low MC plant and one less unit in the high MC plant. Thus SMC for the monopolist across all plants remains the horizontal sum of the SMC curves for individual plants, as in a competitive industry.
 - 6 Cartels are discussed in Chapter 9.

CHAPTER 9

Market structure and imperfect competition

Learning outcomes

By the end of this chapter, you should be able to:

- 1 recognize imperfect competition, oligopoly and monopolistic competition
- 2 understand how cost and demand affect market structure
- 3 describe how globalization changes domestic market structure
- 4 identify equilibrium in monopolistic competition
- 5 recognize the tension between collusion and competition in a cartel
- 6 describe game theory and strategic behaviour
- 7 define the concepts of commitment and credibility
- 8 analyse reaction functions and Nash equilibrium
- 9 describe Cournot and Bertrand competition
- 10 understand the Stackelberg leadership
- 11 recognize why there is no market power in a contestable market
- 12 define innocent and strategic entry barriers

Perfect competition and pure monopoly are useful benchmarks of the extremes of market structure. Most markets are between the extremes. What determines the structure of a particular market? Why are there 10 000 florists but only a few chemical producers? How does the structure of an industry affect the behaviour of its constituent firms?

A perfectly competitive firm faces a horizontal demand curve at the market price. It is a price-taker. Any other type of firm faces a

downward-sloping demand curve for its product and is **imperfectly competitive**.

For a pure monopoly, the demand curve for the firm and the industry coincide. We now distinguish between two intermediate cases of an imperfectly competitive market structure.

The car industry is an **oligopoly**. The price of Volkswagen cars depends not only on Volkswagen's own output and sales but also on the output of other car makers like Ford and Toyota. The corner grocer's shop is a **monopolistic competitor**. Its output is a subtle package of physical goods, personal service and convenience for local customers. It can charge a slightly higher price than an out-of-town supermarket. But, if its prices are too high, even local shoppers travel to the supermarket.

An **imperfectly competitive firm** faces a downward-sloping demand curve. Its output price reflects the quantity of goods it makes and sells.

An **oligopoly** is an industry with few producers, each recognizing their interdependence.

An industry with **monopolistic competition** has many sellers of products that are close substitutes for one another. Each firm has only a limited ability to affect its output price.

As with most definitions, the lines between different market structures can get blurred. One reason is ambiguity about the relevant definition of the market. Is Eurostar a monopoly in cross-channel trains or an oligopolist in cross-channel travel? Similarly, when a country trades in a competitive world market, even the sole domestic producer may have little influence on market price. We can never fully remove these ambiguities, but Table 9.1 shows some things to bear in mind as we proceed through this chapter. The table includes the ease with which new firms can enter the industry, which affects the ability of existing firms to maintain high prices and supernormal profits in the long run.

Table 9.1 Market structure

Competition	Number of firms	Ability to affect price	Entry barriers	Example
Perfect	Lots	Nil	None	Fruit stall

Competition	Number of firms	Ability to affect price	Entry barriers	Example
Imperfect: Monopolistic	Many	Little	Small	Corner shop
Oligopoly	Few	Medium	Bigger	Cars
Monopoly	One	Large	Huge	Post Office

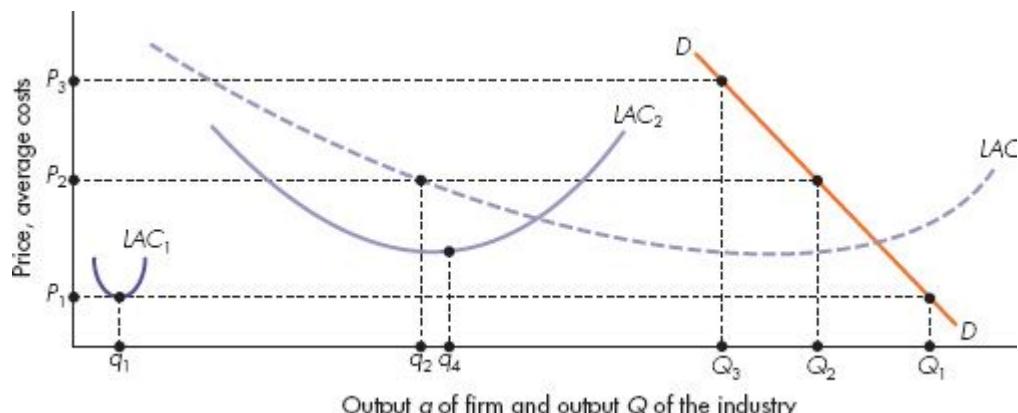
9.1

Why market structures differ

Some industries are legal monopolies, the sole licensed producers. Patent laws may confer temporary monopoly on producers of a new process. Ownership of a raw material may confer monopoly status on a single firm. We now develop a general theory of how demand and cost interact to determine the likely structure of each industry.

The car industry is not an oligopoly one day but perfectly competitive the next. Long-run influences determine market structures.

Figure 9.1 shows the demand curve DD for the output of an industry in the long run. Suppose all firms and potential entrants face the average cost curve LAC_1 . At the price P_1 , free entry and exit means that each firm produces q_1 . With the demand curve DD , industry output is Q_1 . The number of firms in the industry is $N_1 = (Q_1/q_1)$. If at q_1 , the minimum average cost output on LAC_1 is small relative to DD , N_1 will be large. Each firm has a tiny effect on industry supply and market price. We have found a perfectly competitive industry.



DD is the industry demand curve. In a competitive industry, minimum efficient scale occurs at an output level q_1 , when firms have average cost curves LAC_1 . The industry can support a very large number of firms whose total output is Q_1 at the price P_1 . When LAC_3 describes average costs, the industry will be a natural monopoly. When a single firm produces the entire industry output, no other firm can break into the market and make a profit. For intermediate positions such as LAC_2 the industry can support a few firms in the long run, and no single firm can profitably meet the entire demand. The industry will be an oligopoly.

Figure 9.1 Demand, costs and market structure

Next, suppose that each firm has the cost curve LAC_3 . Scale economies are vast relative to the market size. At the lowest point on LAC_3 , output is big relative to the demand curve DD . Suppose initially two firms each make q_2 . Industry output is Q_2 . The market clears at P_2 and both firms break even. If one firm expands a bit, its average costs fall. Its higher output also bids the price down. With lower average costs, that firm survives but the other firm loses money. The firm that expands undercuts its competitor and drives it out of business.

A **natural monopoly** enjoys such scale economies that it has no fear of entry by others.

This industry is a **natural monopoly**. Suppose Q_3 is the output at which its marginal cost and marginal revenue coincide. The price is P_3 and the natural monopoly makes supernormal profits. There is no room in the industry for other firms with access to the same LAC_3 curve.

A new entrant needs a big output to get average costs down. Extra output on this scale so depresses the price that both firms make losses. The potential entrant cannot break in.

Finally, we show the LAC_2 curve with more economies of scale than a competitive industry but fewer than a natural monopoly. This industry supports at least two firms enjoying scale economies near the bottom of their LAC_2 curves. It is an oligopoly. Attempts to expand either firm's output beyond q_4 quickly meet decreasing returns to scale and prevent a firm driving competitors out of business.

Minimum efficient scale is the lowest output at which a firm's LAC curve stops falling.

The crucial determinant of market structure is **minimum efficient scale** relative to the size of the total market as shown by the demand curve. Table 9.2 summarizes our analysis of the interaction of market size and minimum efficient scale. When the demand curve shifts to the left, an industry previously with many firms may have room for only a few. Similarly, a rise in fixed costs, raising the minimum efficient scale, reduces the number of firms. In the 1950s there were many European aircraft makers. Today, the research and development costs of a major commercial airliner are huge. Apart from the co-operative European venture Airbus Industries, only the American giant Boeing survives.

Table 9.2 Demand, cost and market structure

Minimum efficient scale relative to market size		
Tiny	Intermediate	Large
Perfect competition	Oligopoly	Natural monopoly

Monopolistic competition lies between oligopoly and perfect competition. Monopolistic competitors supply different versions of the same product, such as the particular location of a newsagent.

Evidence on market structure

The larger the minimum efficient scale relative to the market size, the fewer the number of plants – and probably the number of firms – in the industry. What number of plants (NP) operating at minimum efficient scale does a market size allow? Chapter 7 discussed estimates of minimum efficient scale in different industries. By looking at the total purchases of a product, we can estimate market size. Hence we can estimate NP for each industry.

Even industries with only a few key players have some small firms on the fringe. The total number of firms can be a misleading indicator of the structure of the industry. In the UK there are many grocery shops; however, four large supermarkets, namely Tesco, Sainsbury's, Morrison and Asda, jointly represent 70 per cent of total grocery retail sales. That industry is therefore quite concentrated, with only few firms serving the

majority of the market. Economists use various indices to measure market concentration. A common measure used is the **N-firm concentration ratio**.¹ This is a measure of the number of key firms in an industry. For example, the three-firm concentration ratio tells us the market share of the largest three firms. If there are three key firms, they will supply most of the market. If the industry is perfectly competitive, the largest three firms will only have a tiny share of industry output and sales.

The **N-firm concentration ratio** is the market share of the largest N firms in the industry.

It would be nice to look at cross-country evidence to see if market structures always obey our theory. If this is to be an independent check, we really need national data before globalization and European integration became important. Table 9.3 examines evidence for the UK, France and Germany for the mid-1970s.

Table 9.3 Concentration and scale economies

Industry	UK		France		Germany	
	CR	NP	CR	NP	CR	NP
Refrigerators	65	1	100	2	72	3
Cigarettes	94	3	100	2	94	3
Refineries	79	8	60	7	47	9
Brewing	47	11	63	5	17	16
Fabrics	28	57	23	57	16	52
Shoes	17	165	13	128	20	197

Note: Concentration ratio *CR* is the percentage market share of the three largest firms; number of plants *NP* is market size divided by minimum efficient scale.

Sources: F. M. Scherer et al., *The economics of multiplant operation* (Harvard University Press, 1975); F. M. Scherer, *Industrial market structure and economic performance* (Rand McNally, 1980).

CR is the three-firm concentration ratio; that is, the market share of the top three firms. *NP* is the number of plants at minimum efficient scale that the market size allows. If our theory of market structure is correct, industries with large-scale economies relative to market size, and thus few plants *NP*, should have a large concentration ratio *CR*. Such

industries should have few key firms. Conversely, where NP is very high, economies of scale are relatively unimportant and the largest three firms should have a much smaller market share. CR should be low.

Table 9.3 confirms that this theory of market structure fits these facts. Industries such as refrigerator and cigarette manufacture had room for few plants operating at minimum efficient scale: these industries had high degrees of concentration. The largest three firms controlled almost the whole market. Scale economies still mattered in industries such as brewing and petroleum refining: the top three firms controlled about half the market. Industries such as shoemaking quickly met rising average cost curves, had room for many factories operating at minimum efficient scale and thus were much closer to competitive industries. The top three firms in shoemaking controlled under one-fifth of the market.

Globalization and multinationals

Table 9.3 showed data before the rise of globalization and multinationals. **Globalization** reflects cheaper transport costs, better information technology and a deliberate policy of reducing cross-country barriers in order to get efficiency gains from large scale and specialization. **Multinationals** sell in many countries at the same time. They may, or may not, also produce in many countries.

Globalization is the closer integration of markets across countries.

Multinationals are firms operating in many countries simultaneously.

CASE 9.1

MARKET STRUCTURE OF THE PC INDUSTRY

The first personal computer was introduced in 1981 by the International Business Machines Corporation (IBM). After that, the PC industry grew swiftly over the years and is now a global business. Countries like Taiwan, Singapore and, more recently, China began to emerge as key players in the PC industry.

According to the data for the first quarter of 2013, the market structure of the PC industry is as shown in the table below.

Company	Market share (%)
HP	15.7
Lenovo	15.3
Dell	11.8
Acer Group	8.1
ASUS	5.7
Others	43.4

The Chinese company Lenovo acquired IBM's PC business in 2005 to accelerate its access to the global market. It is now the second-biggest player in the industry. The five biggest manufacturers control around 57 per cent of the total PC market. Even though the majority of the market is controlled by few firms, the PC industry is still very competitive. The two main characteristics of this industry are:

- *Vertical product differentiation*: this refers to goods that are close substitutes but differ in terms of quality.
- *Fast technological change*: the PC life cycle lasts only four months on average.

Technology adoption is what allows firms to vertically differentiate their products and to gain market share.

Source: IDC press release, 10 April 2013.

Multinationals affect the analysis implied by Figure 9.1 and Table 9.3. They can produce on a large scale somewhere in the world, where production is cheapest, enjoy all the benefits of scale economies, but still sell small quantities in many different markets.

This has three effects. First, it reduces entry barriers in a particular country. A foreign multinational entrant need not achieve a large market share, and therefore need not bid down the price a lot, to achieve scale economies. These now arise because of success in selling globally. Second, small domestic firms, previously sheltered by entry barriers,

now face greater international competition and may not survive. Third, greater competition by low-cost producers leads *initially* to lower profit margins and lower prices.

However, if there are only a few multinationals, they may drive the higher-cost domestic firms out of business but then collude among themselves to raise prices again. Some of the debate about globalization hinges on which of these two outcomes dominates: the initial price fall or a possible subsequent price increase. We will return to this issue shortly when we analyse collusion.

9.2

Monopolistic competition

In London there are more than 5000 restaurants. Every year new restaurants open and some close down, meaning that barriers to entering and exiting the market are relatively low. Even if there is a large number of sellers and no relevant barriers, the restaurant industry is far from being perfectly competitive. The food produced by different restaurants is not perceived to be identical. The products in that industry are imperfect substitutes. The food offered by a Michelin-starred restaurant is normally perceived as different from the food served at McDonald's. Italian food is not a perfect substitute for Indian food. This means that a restaurant may charge a higher price for its food and still retain some of its customers. This could not happen in a perfectly competitive industry. The restaurant industry is an example of a monopolistically competitive industry.

The theory of monopolistic competition envisages a large number of quite small firms so that each firm can neglect the possibility that its own decisions provoke any adjustment in other firms' behaviour. There is free entry to and exit from the industry in the long run. In these respects, the industry resembles *perfect* competition. What distinguishes *monopolistic* competition is that each firm faces a *downward-sloping* demand curve.

Monopolistic competition describes an industry in which each firm can influence its market share to some extent by changing its price relative to that of its competitors. Its demand curve is not horizontal because different firms' products are only limited substitutes, as in the location of local shops. A lower price attracts some customers from another shop but

each shop always has some local customers for whom convenience is more important than a few pence off the price of a jar of coffee.

Monopolistically competitive industries exhibit *product differentiation*. Firms in a monopolistically competitive market produce goods that are perceived by customers as imperfect substitutes. Corner grocers differentiate by location, hairdressers by customer loyalty. The special feature of a particular restaurant or hairdresser lets it charge a slightly different price from other firms in the industry without losing all its customers.

Monopolistic competition requires not merely product differentiation but also limited opportunities for economies of scale. Firms are small. With lots of producers, each can neglect its interdependence with any particular rival. Many examples of monopolistic competition are service industries where economies of scale are small.

The industry demand curve shows the total output demanded at each price if all firms in the industry charge that price. The market share of each firm depends on the price it charges and on the number of firms in the industry. For a given number of firms, a shift in the industry demand curve shifts the demand curve for the output of each firm. For a given industry demand curve, having more (fewer) firms in the industry shifts the demand curve of each firm to the left (right) as its market share falls (rises). But each firm faces a downward-sloping demand curve. This implies that firms in monopolistic competition have market power and they are price-setters. For a given industry demand curve, number of firms and price charged by all other firms, a particular firm can raise its market share a bit by charging a lower price.

CONCEPT 9.1

IT'S NOT WHAT IT LOOKS LIKE

An investor seeking to hold assets in a mutual fund is a consumer with many choices: in 2001, there were 8307 US mutual funds in operation. A mutual fund investor's choice set has also been growing robustly over time: while there were 834 mutual funds in operation in 1980, this nearly quadrupled to 3100 by 1990, and almost tripled again by 2001. So it appears that the mutual fund market in the US is a market with many firms, most of whose

mutual funds are pretty homogeneous, and there is reasonably free entry. Is the market for mutual funds a competitive market?

The answer appears to be no. The fees that investors pay to hold assets in funds are really dispersed, meaning they differ even for mutual funds that are almost homogeneous in their performance. Why should prices be different for goods that are almost homogeneous? The reason is that there are other elements, apart from pure performance of the funds, which can affect investor choice. For example, 60 per cent of investors reported consulting a financial adviser before purchase, implying that the ability of the financial adviser is an important element in investor choice. Funds can have divergent taxable distribution rates for a given return pattern; clearly, investors prefer less tax exposure, all else being equal.

All those facts can explain why products that may appear homogeneous in their physical characteristics can still have some sort of product differentiation that can explain their different prices.

Source: A. Hortaçsu and C. Syverson, ‘Product differentiation, search costs and competition in the mutual fund industry: A case study of S&P index funds’, NBER working paper 9728, 2003.

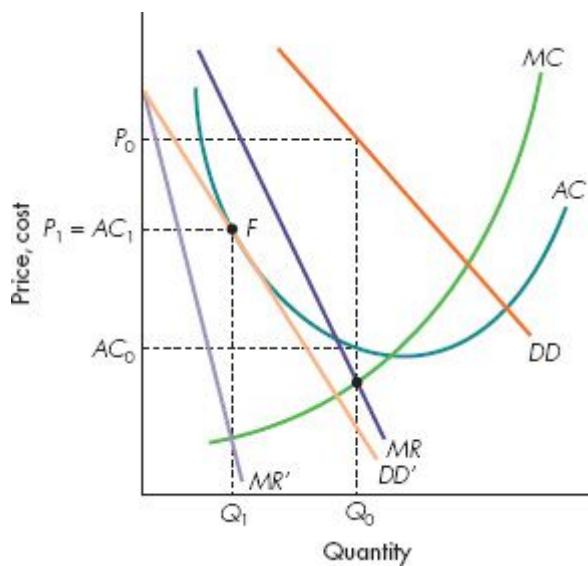
Figure 9.2 shows a firm’s supply decision. Given its demand curve DD and marginal revenue curve MR , the firm makes Q_0 at a price P_0 , making short-run profits $Q_0(P_0 - AC_0)$. In the long run, these profits attract new entrants, diluting the market share of each firm in the industry, shifting their demand curves to the left. Entry stops when each firm’s demand curve shifts so far left that price equals average cost and firms just break even. In Figure 9.2 this occurs when demand is DD' . The firm makes Q_1 at a price P_1 in the **tangency equilibrium** at F .

In monopolistic competition, in the long-run **tangency equilibrium** each firm’s demand curve just touches its AC curve at the output level at which MC equals MR. Each firm maximizes profits but just breaks even. There is no more entry or exit.

Note two things about the firm’s long-run equilibrium at F . First, the firm is *not* producing at minimum average cost. It has excess capacity. It

could reduce average costs by further expansion. However, its marginal revenue would be so low that this is unprofitable. Second, the firm has some monopoly power because of the special feature of its particular brand or location. Price exceeds marginal cost.

This explains why firms are usually eager for new customers prepared to buy additional output at the *existing* price. We are a race of eager sellers and coy buyers. It is purchasing agents who get Christmas presents from sales reps, not the other way round. In contrast, a perfectly competitive firm does not care if another buyer shows up at the existing price. With price equal to marginal cost, the firm is already selling as much as it wants to sell.



In the short run the monopolistic competitor faces the demand curve DD and sets MC equal to MR to produce Q_0 at a price P_0 . Profits are $Q_0 \times (P_0 - AC_0)$. Profits attract new entrants and shift each firm's demand curve to the left. When the demand curve reaches DD' we reach the long-run tangency equilibrium at F. The firm sets MC equal to MR' to produce Q_1 at which P_1 equals AC_1 . Firms are breaking even and there is no further entry.

Figure 9.2 Equilibrium for a monopolistic competitor

9.3

Oligopoly and interdependence

Under perfect competition or monopolistic competition, there are many firms in the industry. Each firm can ignore the effect of its own actions

on rival firms. However, the key to an oligopolistic industry is the need for each firm to consider how its own actions affect the decisions of its relatively few competitors. Each firm has to guess how its rivals will react. Before discussing what constitutes a smart guess, we introduce the basic tension between competition and collusion when firms know that they are interdependent. Initially, for simplicity, we neglect the possibility of entry and focus on existing firms.

The profits from collusion

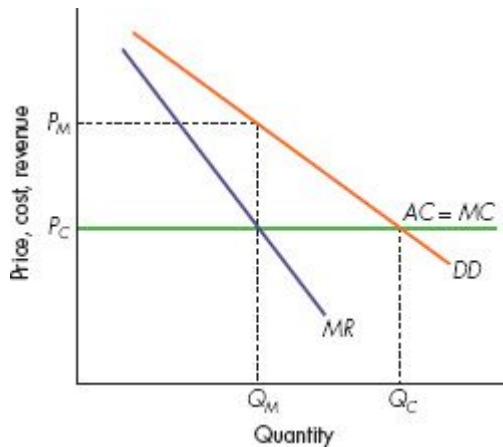
Collusion is an explicit or implicit agreement to avoid competition.

As sole decision maker in the industry, a monopolist would choose industry output to maximize total profits. Hence, the few producers in an industry can maximize their total profit by setting their total output as if they were monopolists.

Figure 9.3 shows an industry where each firm, and the whole industry, has constant average and marginal costs at the level P_C . Chapter 8 showed that a competitive industry produces Q_C at a price P_C but a multiplant monopolist maximizes profits by making Q_M at a price P_M . If the oligopolists collude to produce Q_M , they act as a *collusive monopolist*. Having decided industry output, the firms agree how to share total output and profits among themselves.

However, it is hard to stop firms cheating on the collective agreement. In Figure 9.3 joint profit is maximized at a total output Q_M and price P_M . Yet each firm can expand output at a marginal cost P_C . Any firm can expand output, selling at a little below the agreed price P_M , and make extra profit since its marginal revenue exceeds its marginal cost. This firm gains at the expense of its collusive partners. Industry output is higher than the best output Q_M , so total profits fall and other firms suffer.

Oligopolists are torn between the desire to collude, in order to maximize joint profits, and the desire to compete, in order to raise market share and profits at the expense of rivals. Yet if all firms compete, joint profits are low and no firm does very well. Therein lies the dilemma.



By colluding to restrict industry output Q_M , joint profits are maximized and equal to those which a multi-plant monopolist would obtain. But each firm, with a marginal cost of P_C , has an incentive to cheat on the collusive agreement and to expand its own output.

Figure 9.3 Collusion versus competition

Cartels

Collusion between firms is easiest if formal agreements are legal. Such arrangements, called *cartels*, were common in the late nineteenth century, agreeing market shares and prices in many industries. Cartels are now outlawed in Europe, the US and many other countries. There are big penalties for being caught, but informal agreements and secret deals are sometimes discovered even today.

Cartels across continents are harder to outlaw. The most famous cartel is OPEC, the Organization of Petroleum Exporting Countries. Its members meet regularly to set price and output. Initially, OPEC succeeded in organizing quantity reductions to force up the price of oil. Real OPEC revenues rose 500 per cent between 1973 and 1980. Yet many economists predicted that OPEC, like most cartels, would quickly collapse. Usually, the incentive to cheat is too strong to resist and once somebody breaks ranks others tend to follow. One reason that OPEC was successful for so long was the willingness of Saudi Arabia, the largest oil producer, to restrict its output further when smaller members insisted on expansion.

By 1986 Saudi Arabia was no longer prepared to play by these rules and refused to prop up the price any longer. The oil price collapsed from just under \$30 to \$9 a barrel. During 1987–98, apart from a brief period during the First Gulf War, oil prices fluctuated between \$8 and \$20 a

barrel. Only after 1998 did OPEC recover the cohesion it displayed during 1973–85. The Second Gulf War and continuing uncertainty in the Middle East has continued to restrict supply in any case, also underpinning the high oil prices since 2003.

The kinked demand curve

Collusion is much harder if there are many firms in the industry, if the product is not standardized and if demand and cost conditions are changing rapidly. In the absence of collusion, each firm's demand curve depends on how competitors react. Firms must guess how their rivals will behave.

Suppose that each firm believes that its own price cut will be matched by all other firms in the industry, but that a rise in its own price will not induce a price response from competitors. Figure 9.4 shows the demand curve DD that each firm then believes it faces. At the current price P_0 , the firm makes Q_0 . Suppose the firm raises the price. If competitors do not follow, the firm loses some of its customers, who will now buy from the cheaper competitors. The firm loses market share to other firms. The firm's demand curve is elastic above A at prices above the current price P_0 . However, if each firm believes that if it cuts prices this action will be matched by other firms, market shares are unchanged. Lower prices then induce extra sales rises only because the whole industry moves down the market demand curve as prices fall. The demand curve DD is much less elastic for price cuts from the initial price P_0 . The demand faced by each firm has a kink at point A , reflecting the asymmetric effects of a price change.

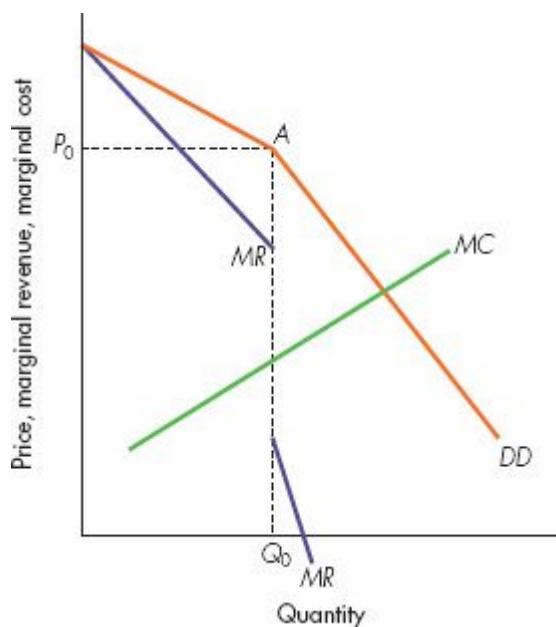
In Figure 9.4 we have to draw marginal revenue MR for each of the separate sections of the kinked demand curve. The firm jumps discontinuously from one part of MR to the other when it reaches the output Q_0 .

Below Q_0 , the elastic part of the demand curve is relevant, and marginal revenue is high since additional output does not depress the price much for existing sales. At the output Q_0 , the firm hits the inelastic portion of its kinked demand curve and marginal revenue becomes much lower: now that demand is less elastic, further output increases require much lower prices to sell the extra output, hitting revenue from existing sales.

Q_0 is the profit-maximizing output for the firm, given its belief about how competitors respond.

Suppose the MC curve of a single firm shifts up or down by a small amount. Since the MR curve has a discontinuous vertical segment at Q_0 , it remains optimal to make Q_0 and charge the price P_0 . In contrast, a monopolist facing a continuously downward-sloping MR curve would adjust quantity and price when the MC curve shifted. The kinked demand curve model may explain the empirical finding that firms do not always adjust prices when costs change.

It does not explain what determines the initial price P_0 . One interpretation is that it is the collusive monopoly price. Each firm believes that an attempt to undercut its rivals will provoke them to cooperate among themselves and retaliate in full. However, its rivals will be happy for it to charge a higher price and see it lose market share.



An oligopolist believes rivals will match price cuts but not price rises. The oligopolist's demand curve is kinked at A. Price rises lead to a large loss of market share, but price cuts increase quantity only by increasing industry sales. Marginal revenue is discontinuous at Q_0 . The oligopolist produces Q_0 , the output at which MC crosses the MR schedule.

Figure 9.4 The kinked demand curve

If we interpret P_0 as the collusive monopoly price, we can contrast the effect of a cost change for a single firm and a cost change for all firms. The latter shifts the marginal cost curve up for the entire industry, raising the collusive monopoly price. Each firm's kinked demand curve shifts up since the monopoly price P_0 has risen. Hence, we can reconcile the stickiness of a firm's price with respect to changes in its own costs alone, and the speed with which the entire industry marks up prices when all firms' costs increase. Examples of the latter are higher taxes on the industry's product or a union wage increase across the whole industry.

9.4

Game theory and interdependent decisions

A good poker player sometimes bluffs. You can win with a bad hand if your opponents misread it for a good hand. Similarly, by having bluffed in the past and been caught, you may persuade opponents to bet a lot when you have a terrific hand.

Like poker players, oligopolists try to anticipate their rivals' moves to determine their own best action. To study interdependent decision making, we use *game theory*. The *players* in the **game** try to maximize their own *payoffs*. In an oligopoly, the firms are the players and their payoffs are their profits in the long run. Each player must choose a strategy. Being a pickpocket is a strategy. Lifting a particular wallet is a move.

A **game** is a situation in which intelligent decisions are necessarily interdependent.

As usual, we are interested in equilibrium. In most games, each player's best **strategy** depends on the strategies chosen by other players. It is silly to be a pickpocket when the police have CCTV cameras or to play four centre backs when the opponents have no proven goal-scorers.

A **strategy** is a game plan describing how a player acts, or moves, in each possible situation.

An equilibrium in a game is a situation where each player plays the best strategy given the strategy of the others. In this case, no player would like to change her chosen strategy. This definition of equilibrium, and its application to game theory, was invented by a Princeton University mathematician, **John Nash**.²

In **Nash equilibrium**, each player chooses the best strategy, given the strategies being followed by other players.

Dominant strategies

Sometimes (but not usually) a player's best strategy is independent of those chosen by others. We begin with an example in which each player has a **dominant strategy**.

A **dominant strategy** is a player's best strategy *whatever* the strategies adopted by rivals.

Figure 9.5 shows a game³ between the only two members of a cartel. Each firm can select a high-output or low-output strategy. In each box of Figure 9.5 the green number shows firm A's profits and the purple number firm B's profits for that output combination.

When both have high output, industry output is high, the price is low and each firm makes a small profit of 1. When each has low output, the outcome resembles collusive monopoly. Prices are high and each firm does better, making a profit of 2. Each firm does best (a profit of 3) when it alone has high output: the other firm's low output helps hold down industry output and keep up the price. In this situation we assume the low-output firm makes a profit of 0.

Now we can see how the game will unfold. Consider firm A's decision. It first thinks what to do if firm B has a high-output strategy. Firm A will thus be in one of the two left-hand boxes of Figure 9.5. Firm A gets a profit of 1 by choosing high but a profit of 0 by choosing low. If firm A thinks firm B will choose high output, firm A prefers high output itself.

		Firm B output	
		High	Low
Firm A output	High	1 1	3 0
	Low	0 3	2 2

The green and purple numbers in each box indicate profits to firms A and B, respectively. Whether B pursues high or low output, A makes more profit going high; so does B, whichever strategy A adopts. In equilibrium both go high. Yet both would make greater profits if both went low!

Figure 9.5 The Prisoner's Dilemma game

But firm A must also think what to do if firm B chooses a low-output strategy. This puts firm A in one of the two right-hand boxes. Firm A *still* prefers high output for itself, which yields a profit of 3 whereas low output yields a profit of only 2. Firm A has a dominant strategy. Whichever strategy B adopts, A does better to choose a high-output strategy.

Firm B also has a dominant strategy to choose high output. If firm B anticipates that firm A will go high, facing a choice of the two boxes in the top row, firm B prefers to go high. If B thinks A will go low, B faces a choice from the two boxes in the bottom row of Figure 9.5, but B still wants to go high. Firm B does better to go high whichever strategy A selects. Both firm A and firm B have a dominant strategy to go high. Equilibrium is the top left-hand box. Each firm gets a profit of 1.

Yet both firms would do better, getting a profit of 2, if they colluded to form a cartel and both produced low – the bottom right-hand box. But neither can risk going low. Suppose firm A goes low. Firm B, comparing the two boxes in the bottom row, will then go high, preferring a profit of 3 to a profit of 2. And firm A will be in trouble; earning a profit of 0 in that event. Firm A can figure all this out in advance, which is why its dominant strategy is to go high.

This shows vividly the tension between collusion and competition. In this example, it appears that the output-restricting cartel will never be formed, since each player can already foresee the overwhelming

incentive for the other to cheat on such an arrangement. How, then, can cartels ever be sustained? One possibility is that there exist binding **commitments**.

A **commitment** is an arrangement, entered into voluntarily, that restricts future actions.

If both players in Figure 9.5 could simultaneously sign an enforceable contract to produce low output, they could achieve the co-operative outcome in the bottom right-hand box, each earning profits of 2. This beats the top left-hand box, which shows the Nash equilibrium of the game when collusion cannot be enforced. Without a binding commitment, neither player can go low because then the other player goes high. Binding commitments, by removing this temptation, let both players go low. Both players gain.

This idea of commitment is important and we shall encounter it many times. Just think of all the human activities that are the subject of legal contracts, a simple commitment simultaneously undertaken by two parties or players.

Although this insight is powerful, its application to oligopoly requires care. Cartels within a country are usually illegal and OPEC is not held together by a contract enforceable in international law. Is there a less formal way in which oligopolists can avoid cheating on the collusive low-output solution to the game? If the game is played only once, this is difficult.

Repeated games

In the real world, the game described above is repeated many times: firms choose output levels day after day. Suppose two players try to collude on low output: each announces a *punishment strategy*. If firm A ever cheats on the low-output agreement, firm B says that it will subsequently react by raising its output. Firm A makes a similar promise.

Suppose the agreement has been in force for some time and both firms have stuck to their low-output deal. Firm A assumes that firm B will go low as usual. Figure 9.5 shows that firm A makes a *temporary* gain today if it cheats and goes high. Instead of staying in the bottom right-hand box with a profit of 2, it can move to the top right-hand box and make 3.

However, from tomorrow onwards, firm B will also go high and firm A can then do no better than continue to go high too, making a profit of 1 for ever more. But if A refuses to cheat today, it can continue to stay in the bottom right-hand box and make 2 for ever. In cheating, A swaps a temporary gain for a permanent reduction in future profits. Thus, punishment strategies can sustain an explicit cartel or implicit collusion even if no formal commitment exists.

It is all very well to promise punishment if the other player cheats. But this will affect the other player's behaviour only if the **threat is credible**.

A **credible threat** is one that, after the fact, is still optimal to carry out.

In the preceding example, once firm A has cheated and gone high, it is then in firm B's interest to go high anyway. Hence a threat to go high if A ever cheats is a credible threat.

These insights shed light on the actual behaviour of OPEC in 1986, when Saudi Arabia dramatically raised its output, leading to a collapse of oil prices. In the 1980s, other members of OPEC had gradually cheated on the low-output agreement, trusting that Saudi Arabia would still produce low to sustain a high price and the cartel's prestige. They hoped Saudi threats to adopt a punishment strategy were empty threats. They were wrong. Figure 9.5 shows that, once the others went high, Saudi Arabia had to go high too.

9.5 Reaction functions

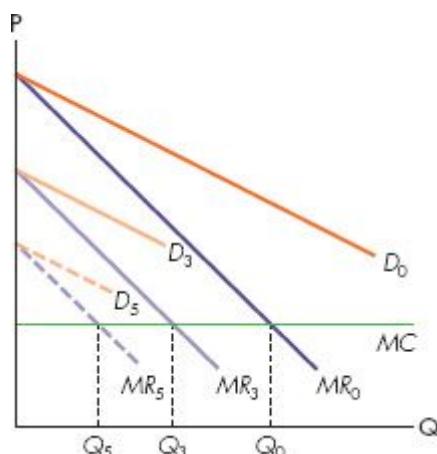
In the previous example, in a one-off game each player had a dominant strategy, to produce high output whatever its rival did. This led to a poor outcome for both players because they were not co-operating despite being interdependent. When the game is repeated, commitments and punishment strategies help players cooperate to find an outcome that is better for both of them.

In punishing a rival, a player's actions change in response to bad behaviour by the rival. Dominant strategies are rare. More usually, each player's best action depends on the actual or expected actions of other players. How a player reacts depends on what it assumes about its rivals'

behaviour. For simplicity we analyse *duopoly*, in which there are only two players.

Cournot behaviour

In 1838 French economist, Augustin Cournot, analysed a simple model of duopoly. Imagine a duopoly in which both firms produce identical products and have the same constant marginal costs MC . Figure 9.6 draws the decision problem for firm A. If firm A assumes that firm B produces 0, firm A gets the whole industry demand curve D_0 . This shows what output firm A can sell given the prices that it charges. From this, firm A calculates the marginal revenue MR_0 , and produces Q_0 to equate its marginal cost and marginal revenue.



Assuming firm B makes 0, firm A faces the market demand curve D_0 and maximizes profits by producing Q_0 to equate marginal cost and marginal revenue. If firm B is assumed to make 3 units, firm A faces the residual demand curve D_3 lying 3 units left of D_0 . Firm A then makes Q_3 . If firm B is assumed to make 5 units, firm A faces D_5 and makes Q_5 . Optimal output for firm A is lower the higher the output that it assumes firm B will make.

Figure 9.6 Cournot behaviour

If, instead, firm A assumes that firm B makes 3 units, firm A faces a demand curve D_3 obtained by shifting the market demand D_0 to the left by 3 units. Firm B gets 3 units and the residual demand is available for firm A. For this demand curve, D_3 , firm A computes the marginal

revenue curve MR_3 and chooses output Q_3 to equate marginal cost and marginal revenue.

Similarly, if firm A expects firm B to make 5 units, firm A shifts D_0 to the left by 5 units to get D_5 , and produces Q_5 in order to equate marginal cost and its marginal revenue MR_5 . The larger the output that firm 2 is expected to make and sell, the smaller the optimal output of firm A. Q_5 is smaller than Q_3 , which is smaller than Q_0 .

Repeating this exercise for every possible belief that firm A has about the output of firm B yields the **reaction function** of firm A.

In the **Cournot model**, a rival's action is its output choice. Figure 9.7 shows the two outputs Q^A and Q^B . From Figure 9.6 firm A makes less the more it thinks that firm B will make. In Figure 9.7 firm A's optimal output choice is the reaction function R^A . If firm B is expected to produce 1 unit less, firm A chooses to raise output by less than 1 unit. This ensures total output falls, thus raising the price. Because this lets firm A earn more on its previous output units, it is not worth raising its output by as much as it expects the output of B to fall. Equivalently, in Figure 9.6 firm A's demand curve shifts more than its marginal revenue curve; hence its rise in output is smaller than the conjectured fall in the output of firm B.

A firm's **reaction function** shows how its optimal output varies with each possible action by its rival.

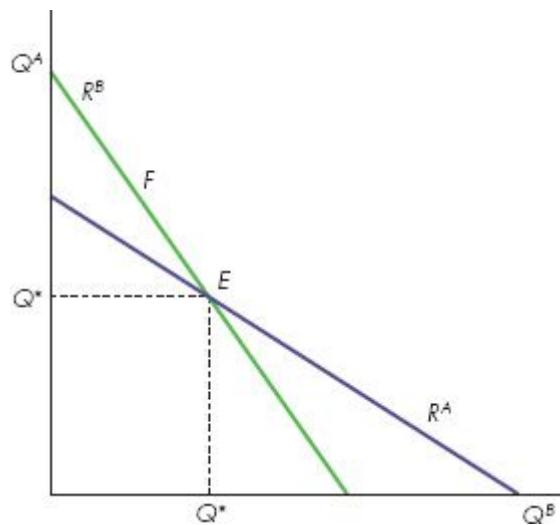
In the **Cournot model**, each firm treats the output of the other firm as given.

In the duopoly, both firms are the same since they have the same marginal cost and produce the same good. Hence firm B faces a similar problem. It makes guesses about the output of firm A, calculates the residual demand curve for firm B, and chooses its best output. Figure 9.7 shows the reaction function R^B for firm B, which also makes less the more that it assumes its rival will produce.

Along each reaction function, each firm makes its best response to the assumed output of the other firm. Only in equilibrium is it optimal for the other firm actually to behave in the way that has been assumed. In

Nash equilibrium, neither firm wishes to alter its behaviour even after its conjecture about the other firm's output is then confirmed.

Since both firms face the same industry demand curve, their reaction functions are symmetric if they also face the same marginal cost curves in Figure 9.6. The two firms then produce the same output Q^* , as shown in Figure 9.7. If costs differed, we could still construct (different) reaction functions and their intersection would no longer imply equal market shares.



R^A is the reaction function of firm A, showing how its optimal output varies with the output it assumes firm B will make and sell. Since firms are similar, R^B is the similar reaction function for firm B, showing its best output given the assumed output by firm A. With these Cournot assumptions about its rival's behaviour, point E is the Nash equilibrium. Each firm's guess about its rival's behaviour is then correct, and neither firm wishes to change its behaviour. If the firms are identical, their reaction functions are asymmetric, and both make the same output Q^* .

Figure 9.7 Nash–Cournot equilibrium

Suppose the marginal cost curve of firm A now shifts down in Figure 9.6. At each output assumed for firm B, firm A now makes more. It moves further down any MR schedule before meeting MC . Hence, in Figure 9.7 the reaction function R^A shifts up, showing firm A makes more output Q^A at any assumed output Q^B of its rival. The new intersection of the reaction functions, say at point F, shows what happens to Nash equilibrium in the Cournot model.

It is no surprise that the output of firm A rises. Why does the output of firm B fall? With lower marginal costs, firm A is optimally making more. Unless firm B cuts its output, the price will fall a lot. Firm B prefers to cut output a little, in order to prop up the price a bit, preventing a big revenue loss on its existing units.

As in our discussion of the Prisoner's Dilemma game in Section 9.4, the Nash–Cournot equilibrium does not maximize the joint payoffs of the two players. They fail to achieve the total output that maximizes joint profits. By treating the output of the rival as given, each firm expands too much. Higher output bids down prices for everybody. In neglecting the fact that its own expansion hurts its rival, each firm's output is too high.

Each firm's behaviour is correct given its assumption that its rival's output is fixed. But expansion by one firm induces the rival to alter its behaviour. A joint monopolist would take that into account and make more total profit.

This is considered in Figure 9.8. Suppose there are two identical firms producing cars. The firms have two possible strategies: co-operate and form a cartel or do not co-operate and compete in quantities. The game is played simultaneously and only once, so it is a one-shot game. If they cooperate (collude), they can set the monopoly price and both obtain half of the monopoly profits. If they compete, they both obtain the Cournot profits, which are lower than in the case of collusion. If a firm is co-operating while the rival deviates from the collusive agreement, the firm deviating steals most of the market and obtains high profits. The other firm receives low profits.

		Firm B	
		Co-operate	Not co-operate
Firm A	Co-operate	10, 10	2, 15
	Not co-operate	15, 2	5, 5

Figure 9.8 Cournot competition and the Prisoner's Dilemma

From Figure 9.8 we can see that firm A has a dominant strategy (to not co-operate), since that strategy, independently of what the rival is doing,

will provide a payoff of 15 or 5 (co-operating will give firm A payoffs of 10 or 2). For firm B, we have a dominant strategy as well. Firm B will always choose not to co-operate. The only Nash equilibrium of the game is for both firms not to co-operate. At that equilibrium, the firms will get profits of 5, lower than in the case of both co-operating.

In this case, firms do not co-operate because the incentive to deviate from the collusive agreement is large. By recognizing that, both firms will simply not co-operate and we are back to the Prisoner's Dilemma case.

MATHS 9.1

DERIVING THE REACTION FUNCTIONS IN A COURNOT DUOPOLY

Consider a market in which there are two firms, A and B, producing the same good and competing on quantities. The inverse market demand is given by $P = a - bQ$, where $Q = Q_A + Q_B$ is the total quantity produced in the market and is simply the sum of what is produced by firm A and firm B. Assume that the cost functions of the two firms are the same, meaning that the two firms are equal (the case of an unequal cost function can be handled easily too).

The cost function of firm A is $TC_A = cQ_A$, while for firm B it is $TC_B = cQ_B$, where $c > 0$ is the marginal cost.

The reaction function (or best response) of firm A tells us how the output produced by firm A depends on the output produced by firm B. The reaction function for firm B is defined in a similar way.

Each firm maximizes profits. This means that each firm chooses a level of output such that the marginal revenue of selling that output is equal to the marginal cost of producing it. The total revenue for firm A is given by $TR_A = P \times Q_A$. Using the inverse demand for substituting for P , we get $TR_A = [a - b(Q_A + Q_B)] \times Q_A = aQ_A - bQ_A^2 - bQ_A Q_B$. As we can see, the total revenue of firm A now depends on the output chosen by firm B as well (Q_B). For firm B, using a similar argument, the total revenue is $TR_B = aQ_B - bQ_B^2 - bQ_B Q_A$.

The marginal revenue functions for the two firms are:

$$MR_A \equiv \frac{\partial TR_A}{\partial Q_A} = a - 2bQ_A - bQ_B \quad \text{and} \quad MR_B \equiv \frac{\partial TR_B}{\partial Q_B} = a - 2bQ_B - bQ_A$$

The marginal costs of the firms are the same:

$$MC_A \equiv \frac{dTC_A}{dQ_A} = c \quad \text{and} \quad MC_B \equiv \frac{dTC_B}{dQ_B} = c$$

The reaction function of each firm is found for the profit-maximization conditions:

$$MR_A = MC_A \quad \text{and} \quad MR_B = MC_B$$

Using our data to express those two conditions, we have:

$$a - 2bQ_A - bQ_B = c \quad \text{and} \quad a - 2bQ_B - bQ_A = c$$

From those two conditions we can find the reaction functions of each firm. For firm A, the reaction function is:

$$Q_A = \frac{a - c}{2b} - \frac{Q_B}{2} \tag{1}$$

For firm B, we have:

$$Q_B = \frac{a - c}{2b} - \frac{Q_A}{2} \tag{2}$$

Notice that the reaction function of each firm depends negatively on the output produced by the rival. If firm B increases its output level, the best response by firm A is to reduce its output level. Reaction function (1) tells us the output that maximizes the profits of firm A, given the output chosen by firm B.

The Nash equilibrium of the Cournot model is where the two reaction functions above are mutually consistent, meaning they cross.

Therefore, we just need to solve a system of two equations in two variables Q_A, Q_B .

By solving the system of equations (1) and (2), we have:

$$Q_A = \frac{a - c}{3b} \quad \text{and} \quad Q_B = \frac{a - c}{3b}$$

Since the two firms are identical by assumption, they must produce the same level of output. The total output produced in the market is therefore: $Q_A + Q_B = 2[(a - c)/3b]$. The equilibrium price is found through the inverse demand function:

$$P = \frac{a + 2c}{3}$$

Bertrand behaviour

To show how the assumption about rivals' behaviour affects reaction functions and hence Nash equilibrium, consider a different model suggested by another French economist, Joseph Bertrand.

Each firm produces the same good. In contrast to Cournot competition, each firm now decides a price (and hence an output) reflecting the price it expects its rival to set. We could go through a similar analysis to the Cournot model, find reaction curves showing how the *price* set by each firm depends on the *price* set by its rival, and hence find the Nash equilibrium in prices for the **Bertrand model**. Knowing the equilibrium price, we could work out equilibrium quantity. If the firms are identical, again they divide the market equally. However, in the Bertrand model, it is easy to see what the Nash equilibrium must be. It is the perfectly competitive outcome: price equals marginal cost. How do we know?

In the **Bertrand model** of oligopoly, each firm treats the *prices* of rivals as given.

Suppose firm B sets a price above its marginal cost. Firm A can grab the whole market by setting a price a little below that of firm B. Since firm B can anticipate this, it must set a lower price. This argument keeps working until, in Nash equilibrium, both firms price at marginal cost and split the market between them. There is then no incentive to alter behaviour.

Comparing Bertrand and Cournot

Under Bertrand behaviour, Nash equilibrium entails price equal to marginal cost, so industry output is the same as in a perfectly competitive market. Under Cournot behaviour, Nash equilibrium entails lower industry output and a higher price than under Bertrand behaviour. Under Cournot behaviour each firm makes positive profits. But the firms do not co-operate. A joint monopolist would make more profit by coordinating output decisions. Industry output would be even lower and the price even higher.

Thus, Nash equilibrium depends on the *particular* assumption each firm makes about its rival's behaviour. Generally, economists prefer the Cournot model. In practice, few oligopolies behave like a perfectly competitive industry, as the Bertrand model predicts.

Moreover, since prices can be changed rapidly, treating a rival's *price* as fixed does not seem plausible. In contrast, we can interpret the Cournot model as saying that firms first choose *output capacity* and then set price. Since capacity takes time to alter, this makes more sense.

CONCEPT 9.2

MERGERS AND COMPETITION POLICY

Two firms can unite in two different ways: via a takeover bid or a merger. When a firm makes a takeover bid, managers of the 'victim' firm usually resist since they are likely to lose their jobs, but the shareholders will accept if the offer is sufficiently attractive.

From now on we use mergers as shorthand for both forms of union. Mergers can be distinguished in the following way: a *horizontal merger* is the union of two firms at the same production stage in the same industry. A *vertical merger* is the union of two firms at different production stages in the same industry. In a *conglomerate merger*, the production activities of the two firms are unrelated.

Are mergers in the public interest, or do they just create private monopolies?

On the one hand, a merger reduces the number of competitors in a market. Consider a market with six main competitors. If two of them merge, the number of competitors is reduced to five. This reduction in competition is beneficial for all the remaining firms in the market, but it may be detrimental for consumers. Less

competition may result in higher prices. The merger of two large firms gives them market power from a large market share. The merged company is likely to restrict output and increase prices – a deadweight burden for society as a whole.

On the other hand, two firms may merge for efficiency reasons. The new firm may be more efficient than the two separate firms; there may be gains to co-ordination and planning and in managerial and financial aspects. If companies achieve any of these benefits, they will increase productivity and lower costs. Competition policy related to mergers must compare the gains (potential cost reduction) with the costs (larger market power).

The table below shows annual averages of takeovers and mergers involving UK firms. It shows dramatic merger booms in the late 1980s and late 1990s, which coincided with high stock market values, which raised the value of both firms involved in the merger.

The proliferation of large companies through merger would not have been possible if there had been a tough anti-merger policy.

There are currently two grounds for referring a prospective merger to an investigation by the Competition Commission: (1) that the merger will promote a new monopoly as defined by the 25 per cent market share used in deciding references for existing monopoly positions; or (2) that the company taken over has an annual UK turnover of at least £70 million.

UK takeovers and mergers, 1986–2012 (annual averages)

	Number	Value (1998 £bn)
1986–89	1300	43
1990–98	585	20
1999–00	540	61
2001–06	600	30
2007–12	650	47

Sources: *British Business Trends*, 1989; *Business Trends*, 1997; ONS.

Since the merger legislation was introduced in 1965, only 4 per cent of all merger proposals have been referred to the Competition Commission. For much of the period, government policy has been to consent to, or actively encourage, mergers. In believing that the

benefits would outweigh the costs, UK merger policy reflected two assumptions. The first was that the cost savings from economies of scale and more intensive use of scarce management talent could be quite large. The second was that the UK was part of an increasingly competitive world market so that the monopoly power of the merged firms, and the corresponding social cost of the deadweight burden, would be small. Large as they were, the merged firms were small in relation to European or world markets, and would face relatively elastic demand curves, giving little scope to raise price above marginal cost.

Finally, as with competition policy, EU legislation takes precedence where this is appropriate. It is not appropriate in assessing whether a merger of two UK supermarkets should be allowed, since this predominantly affects only UK consumers. However, the European Commission will investigate mergers involving enterprises with an aggregate worldwide annual turnover of over €5 billion and where the aggregate EU-wide turnover of each of the enterprises exceeds €250 million.

First-mover advantage and the Stackelberg model

So far we have assumed that the two duopolists make decisions simultaneously. Suppose one firm can choose output before the other. This means that we move from a simultaneous game to a sequential game structure. Does it help to move first?

To anticipate how firm B behaves once the output of firm A is fixed, firm A examines the reaction function of firm B as derived in Figures 9.6 and 9.7. In setting output, firm A then takes account of how its own output decisions *affect* output by firm B.

Firm A thus has a different reaction function. Figure 9.7 showed the Cournot reaction function R^A , treating Q^B as chosen independently of Q^A . Now firm A uses the reaction function RB to deduce that a higher output Q^A induces a *lower* output Q^B . Hence, firm A expects its own output expansion to bid the price down *less* than under Cournot behaviour. Its marginal revenue schedule is higher up. Firm A knows that firm B will help prop up the price by cutting Q^B in response to a rise in Q^A .

Facing a higher MR schedule as a **Stackelberg** leader than under Cournot behaviour, firm A produces more than under Cournot behaviour. Firm B makes less because it must react to the fact that a high output Q^A is already a done deal. Firm A ends up with higher output and profits than under Cournot behaviour but firm B has lower output and lower profit. Firm A has a **first-mover advantage**.

In The **Stackelberg model**, firm B can observe the output already fixed by firm A. In choosing output, firm A must thus anticipate the subsequent reaction of firm B.

A **first-mover advantage** means that the player moving first achieves higher payoffs than when decisions are simultaneous.

Moving first acts like a commitment that prevents your subsequent manipulation by the other player. Once firm A has built a large output capacity, firm B has to live with the reality that firm A will produce large output. The best response of firm B is then low output. Raising up the output price helps firm A. Being smart, firm A had already figured all that out.

In some industries, firms are fairly symmetric and Cournot behaviour is a good description of how these oligopolists behave. Other industries have a dominant firm, perhaps because of a technical edge or privileged location. That firm may be able to act as a Stackelberg leader and anticipate how its smaller rivals will then react.

MATHS 9.2

THE STACKELBERG MODEL

Consider a market in which operate two firms, A and B. Firm A is the leader while firm B is the follower.

The market inverse demand function is $P = a - b(Q_A + Q_B)$. Assume that the total cost of firm A is $TC_A = c_A Q_A$, while for firm B it is $TC_B = c_B Q_B$. The marginal cost of firm A is therefore c_A and for firm B is c_B . The follower (firm B) takes the output produced by the leader (firm A) as given. Let's look at the behaviour of the follower first. The total revenue function of firm B is $TR_B = aQ_B -$

$bQ_A Q_B - bQ_B^2$. The marginal revenue of the follower is $MR_B = a - bQ_A - 2bQ_B$. The reaction function of the follower comes from the profit-maximizing condition of firm B, $MR_B = MC_B$. Using our data, that condition implies $a - bQ_A - 2bQ_B = c_B$. Solve for Q_B :

$$Q_B = \frac{a - c_B}{2b} - \frac{Q_A}{2} \quad (1)$$

This is the reaction function of firm B. Firm A is the leader and takes into account that the follower has the reaction function given by (1).

The total revenue function of the leader is therefore:

$$TR_A = \left[a - b\left(Q_A + \frac{a - c_B}{2b} - \frac{Q_A}{2}\right) \right] \times Q_A \quad (2)$$

The term in the square brackets is just the inverse demand (and so the price) once we take into account the reaction function of the follower. Equation (2) becomes:

$$TR_A = aQ_A - bQ_A^2 - \frac{a - c_B}{2}Q_A + \frac{b}{2}Q_A^2$$

The marginal revenue of the leader is therefore:

$$MR_A \equiv \frac{\partial TR_A}{\partial Q_A} = a - 2bQ_A + bQ_A \frac{a - c_B}{2}$$

The output that maximizes the profits of firm A comes from the condition $MR_A = MC_A$, that is:

$a - 2bQ_A + bQ_A = (a - c_B)/2 = c_A$. Solve for Q_A :

$$Q_A = \frac{a + c_B - 2c_A}{2b}$$

Once we know the optimal choice for the leader, we can go back to the reaction function of the follower and substitute for the Q_A we just found:

$$Q_B = \frac{a - c_B}{2b} - \left(\frac{a + c_B - 2c_A}{4b} \right)$$

Simplifying that equation we obtain:

$$Q_B = \frac{a - 3c_B + 2c_A}{4b}$$

9.6 ENTRY AND POTENTIAL COMPETITION

So far we have discussed imperfect competition between existing firms. To complete our understanding of such markets, we must also think about the effect of potential competition from new entrants to the industry on the behaviour of existing or incumbent firms. Three cases must be distinguished: where entry is completely easy, where it is difficult by accident and where it is difficult by design.

Contestable markets

Free entry to, and exit from, the industry is a key feature of perfect competition, a market structure in which each firm is tiny relative to the industry. Suppose, however, that we observe an industry with few incumbent firms. Before assuming that our previous analysis of oligopoly is needed, we must think hard about entry and exit. The industry may be a **contestable market**.

A **contestable market** has free entry and free exit.

By free entry, we mean that all firms, including both incumbents and potential entrants, have access to the same technology and hence have the same cost curves. By free exit, we mean that there are no *sunk* or irrecoverable costs: on leaving the industry, a firm can fully recoup its previous investment expenditure, including money spent on building up knowledge and goodwill.

A contestable market allows *hit-and-run* entry. If the incumbent firms, however few, do not behave as if they were a perfectly competitive industry ($p = MC = \text{minimum } LAC$), an entrant can step in, undercut them and make a temporary profit before quitting again.

As globalization proceeds, we should remember that foreign suppliers are important potential entrants. This can take two forms. First, if monopoly profits are too high in the domestic market, competition from imports may augment supply, bidding down prices and profits in the domestic market. In the extreme case, in which imports surge whenever domestic prices rise above the world price, we are back in the competitive world analysed in Chapter 8.

Globalization also raises the likelihood that foreign firms will set up production facilities in the home market, a tangible form of entry. By raising the supply of potential entrants, globalization increases the relevance of contestable markets as a description of market structure. Moreover, we normally think of an entrant as having to start from scratch. When an existing foreign firm enters the domestic market, its production and marketing expertise may already be highly developed.

Globalization may be a two-edged sword. On the one hand, it raises the size of the relevant market and makes entry easier. On the other hand, by allowing multinationals to become vast by operating in many countries simultaneously, globalization may encourage the formation of large firms that then have substantial market power wherever they operate. Coke and Pepsi are slugging it out for global dominance and Virgin Cola provides only limited competition, even in the UK.

The theory of contestable markets remains controversial. There are many industries in which sunk costs are hard to recover or where the initial expertise may take an entrant some time to acquire, placing it at a temporary disadvantage against incumbent firms. Nor, as we shall shortly see, is it safe to assume that incumbents will not change their behaviour when threatened by entry. But the theory does vividly illustrate that market structure and incumbent behaviour cannot be deduced simply by counting the number of firms in the industry.

In the previous chapter, we were careful to stress that a monopolist is a sole producer *who can completely discount fear of entry*. We now refine the classification in Table 9.1 by discussing entry in more detail.

Innocent entry barriers

Our discussion of entry barriers distinguishes those that occur anyway and those that are deliberately erected by incumbent firms.

The American economist, Joe Bain, identified three types of entry barrier: product differentiation, absolute cost advantages and scale economies. The first of these is not an **innocent barrier**, as we shall shortly explain. Absolute cost advantages, where incumbent firms have lower cost curves than those that entrants will face, may be innocent. If it takes time to learn the business, incumbents will face lower costs, at least in the short run. If they are smart, they may already have located in the most advantageous site. In contrast, if incumbents have undertaken investment or R&D specifically with a view to deterring entrants, this is not an innocent barrier. We take up this issue shortly.

An **innocent entry barrier** is one not deliberately erected by incumbent firms.

Figure 9.1 showed the role of scale economies as an innocent entry barrier. If minimum efficient scale is large relative to the industry demand curve, an entrant cannot get into the industry without considerably depressing the market price, and it may prove simply impossible to break in at a profit.

The greater such innocent entry barriers, the more appropriate it is to neglect potential competition from entrants. The oligopoly game then comes down to competition between incumbent firms along the lines we discussed in the previous section. Where innocent entry barriers are low, one of two things may happen. Either incumbent firms accept this situation, in which case competition from potential entrants will prevent incumbent firms from exercising much market power – the outcome will be closer to that of perfect competition – or else incumbent firms will try to design some entry barriers of their own.

9.7 Strategic entry deterrence

A *strategy* is a game plan where decision making is interdependent. The word ‘strategic’ is used in everyday language but it has a precise meaning in economics.

In Figure 9.9 a single incumbent firm plays a game against a potential entrant. The entrant can come in or stay out. If the entrant comes in, the incumbent can opt for the easy life, accept the new rival and agree to

share the market – or it can fight. Each party undertakes a series of **strategic moves**. Fighting entry means producing at least as much as before, and perhaps considerably more than before, so that the industry price collapses. In this *price war*, sometimes called *predatory pricing* by the incumbent, both firms do badly and make losses. The top row of boxes in Figure 9.9 shows the profits to the incumbent (in blue) and the entrant (in orange) in each of the three possible outcomes.

A **strategic move** is one that influences the other person's choice, in a manner favourable to oneself, by affecting the other person's expectations of how one will behave.

If the incumbent is unchallenged it does very well, making profits of 5. The entrant of course makes nothing. If they share the market, both make small profits of 1. In a price war, both make losses. How should the game go?

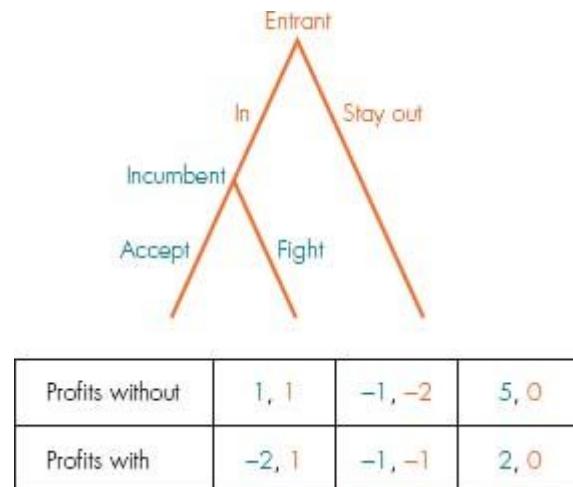


Figure 9.9 Strategic entry deterrence

In the absence of deterrence, if the entrant enters, the incumbent does better to accept entry than to fight. The entrant knows this and enters. Equilibrium is the top left-hand box, and both firms make a profit of 1. But if the incumbent pre-commits an expenditure of 3 which is recouped only if there is a fight, the incumbent resists entry, the entrant stays out and equilibrium is the bottom right-hand box. The incumbent does better, making a profit of 2.

Suppose the entrant comes in. Comparing the left and middle boxes of the top row, the incumbent does better to cave in than to fight. The

entrant can figure this out. Any threat by the incumbent to resist entry is not a credible threat – when it comes to the crunch, it will be better to cave in. Much as the incumbent would like the entrant to stay out, in which case the incumbent would make profits of 5, the equilibrium of the game is that the entrant will come in and the incumbent will not resist. Both make profits of 1, the top left-hand box.

The incumbent, however, may have got its act together before the potential entrant appears on the scene. It may be able to invent a binding pre-commitment, forcing itself to resist entry and thereby scare off a future challenge. The incumbent would be ecstatic if a Martian appeared and guaranteed to shoot the incumbent's directors if they ever allowed an entry to be unchallenged. Entrants would expect a fight, would anticipate a loss of 1, and would stay out, leaving the incumbent with a permanent profit of 5.

In the absence of Martians, the incumbent can achieve the same effect by economic means. Suppose the incumbent invests in expensive spare capacity that is unused at low output. The incumbent has low output in the absence of entry or if an entrant is accommodated without a fight. Suppose in these situations the incumbent loses 3 by carrying this excess capacity. The second row of boxes in Figure 9.9 reduces the incumbent's profits by 3 in these two outcomes. In a price war, however, the incumbent's output is high and the spare capacity is no longer wasted; hence we do not need to reduce the incumbent's profit in the middle column of boxes in Figure 9.9. Now consider the game again.

If the entrant comes in, the incumbent loses 2 by caving in but only 1 by fighting. Hence entry is resisted. Foreseeing this, the entrant does not enter, since the entrant loses money in a price war. Hence the equilibrium of the game is the bottom right-hand box and no entry takes place. **Strategic entry deterrence** has been successful. It has also been profitable. Even allowing for the cost of 3 of carrying the spare capacity, the incumbent still makes a profit of 2, which is better than the profit of 1 in the top left-hand box when no deterrence was attempted and the entrant came in.

The **Strategic entry deterrence** is behaviour by incumbent firms to make entry less likely.

Does deterrence always work? No. Suppose in Figure 9.9 we change the right-hand column. In the top row the incumbent gets a profit of 3 if no entry occurs. Without the pre-commitment, the equilibrium is the top left-hand box, as before. But if the incumbent has to spend 3 on a spare capacity pre-commitment, it now makes a profit of 0 in the bottom right-hand box when entry is deterred. The entrant is still deterred but the incumbent would have done better not to invest in spare capacity and to let the entrant in.

This model suggests that price wars should never happen. If the incumbent really is going to fight, then the entrant should not have entered. This of course requires the entrant to know accurately the profits of the incumbent in the different boxes and therefore correctly predict its behaviour. In the real world, entrants sometimes get it wrong. Moreover, if the entrant has much better financial backing than the incumbent, a price war may be a good investment for the entrant. The incumbent will exit first and thereafter the entrant will be able to cash in and get its losses back with interest.

CASE 9.2

BARRIERS AT THE CHECKOUT

In 2004 the Morrisons supermarket chain finally completed its takeover of rival Safeway. At a stroke, Morrisons was catapulted from the supermarket minnow, with a 6 per cent market share, to a big league player with 17 per cent of the UK market, only marginally less than Sainsbury's, onetime leader of the supermarket industry.

The takeover of Safeway was contested, with Tesco, Asda and Sainsbury's all mounting rival bids to Morrisons'. At one stage, Philip Green, the owner of high-street retailer British Home Stores (Bhs), also registered an interest in Safeway. Safeway was such an attractive target because it provided the last chance to enter the supermarket industry. Without access to land, and facing difficulty getting planning permission for new supermarkets, the only entry mode was a takeover. With Safeway now in the hands of Morrisons, and the industry consolidated into large players, the next takeover will be even more difficult.



Morrisons supermarket in Newport, Isle of Wight. © Editor5807

Is spare capacity the only pre-commitment available to incumbents? Pre-commitments must be irreversible, otherwise they are an empty threat, and they must increase the chances that the incumbent will fight. Anything with the character of fixed and sunk costs may work: fixed costs artificially increase scale economies and make the incumbent keener on high output, and sunk costs cannot be reversed. Advertising to invest in goodwill and brand loyalty is a good example. So is product proliferation. If the incumbent has only one brand, an entrant may hope to break in with a different brand. But if the incumbent has a complete range of brands or models, an entrant will have to compete across the whole product range.

9.8 Summing up

Few industries in the real world are like the textbook extremes of perfect competition and pure monopoly. Most are imperfectly competitive. This chapter introduced you to types of imperfect competition. Game theory in general, and concepts such as commitment, credibility and deterrence, allow economists to analyse many of the practical concerns of big business.

What have we learned? First, market structure and the behaviour of incumbent firms are determined *simultaneously*. Economists used to start with a market structure, determined by the extent of scale economies relative to the industry demand curve, then deduce how the incumbent firms would behave (monopoly, oligopoly, perfect competition), then check out these predictions against performance indicators, such as the extent to which prices exceeded marginal cost. Now we realize that

strategic behaviour by incumbent firms can affect entry, and hence market structure, except where entry is almost trivially easy.

Second, and related, we have learned the importance of *potential* competition, which may come from domestic firms considering entry or from imports from abroad. The number of firms observed in the industry today conveys little information about the extent of the market power they truly exercise. If entry is easy, even a single incumbent or apparent monopolist may find it unprofitable to depart significantly from perfectly competitive behaviour.

Finally, we have seen how many business practices of the real world – price wars, advertising, brand proliferation, excess capacity or excessive research and development – can be understood as strategic competition in which, to be effective, threats must be made credible by prior commitments.

Summary

- **Imperfect competition** exists when individual firms believe they face downward-sloping demand curves. The most important forms are monopolistic competition, oligopoly and pure monopoly.
- **Pure monopoly** status can be conferred by legislation, as when an industry is nationalized or a temporary patent is awarded. When **minimum efficient scale** is very large relative to the industry demand curve, this innocent entry barrier may be sufficiently high to produce a natural monopoly in which all threat of entry can be ignored.
- At the opposite extreme, entry and exit may be costless. The market is **contestable**, and incumbent firms must mimic perfectly competitive behaviour to avoid being flooded by entrants. With an intermediate size of entry barrier, the industry may be an oligopoly.
- **Monopolistic competitors** face free entry to and exit from the industry but are individually small and make similar though not identical products. Each has limited monopoly power in its special brand. In

long-run equilibrium, price equals average cost but exceeds marginal revenue and marginal cost at the tangency equilibrium.

- **Oligopolists** face tension between collusion to maximize joint profits and competition for a larger share of smaller joint profits. **Collusion** may be formal, as in a cartel, or informal. Without **credible threats** of punishment by its partners, each firm faces a temptation to cheat.
- **Game theory** analyses interdependent decisions in which each player chooses a strategy. In the Prisoner's Dilemma game, each firm has a dominant strategy. With binding commitments, both players could do better by guaranteeing not to cheat on the collusive solution.
- A **reaction function** shows one player's best response to the actions of other players. In **Nash equilibrium** reaction functions intersect. No player then wishes to change her decision.
- In **Cournot behaviour** each firm treats the output of its rival as given. In **Bertrand behaviour** each firm treats the price of its rival as given. Nash–Bertrand equilibrium entails pricing at marginal cost. Nash–Cournot equilibrium entails lower output, higher prices and profits. However, firms still fail to maximize joint profits because each neglects the fact that its output expansion hurts its rivals.
- A firm with a **first-mover advantage** acts as a **Stackelberg leader**. By deducing the subsequent reaction of its rival, it produces higher output, knowing the rival will then have to produce lower output. Moving first is a useful commitment.
- **Innocent entry barriers** are made by nature, and arise from scale economies or absolute cost advantages of incumbent firms. **Strategic entry barriers** are made in boardrooms and arise from credible commitments to resist entry if challenged. Only in certain circumstances is strategic entry deterrence profitable for incumbents.

Review questions

EASY

- 1 ‘Since a firm’s optimal behaviour depends on how it believes that its rival(s) will react, there are as many output decisions, and hence equilibriums, as there are guesses about what rivals will do.’ How do economists try to narrow down the assumptions that firms make about their rivals?
- 2 Many of the interesting games are games against the government. Think of a European airline, until recently state-owned, now private but losing money under the pressure of high oil prices and the growth of low-cost airlines. Believing that the government will bail it out if the worse comes to the worst, the airline has no incentive to take today the tough measures needed to make its business profitable. How can the government signal that it will not bail out the airline, forcing the airline to improve or go bust?
- 3 A good-natured parent knows that children sometimes need punishing but also knows that, when it comes to the crunch, the child will be let off with a warning. Can the parent undertake any pre-commitment to make the threat of punishment credible?
- 4 Think of five adverts on television. Is their function primarily informative or the erection of entry barriers to the industry?
- 5 **Common fallacies** Why are these statements wrong? (a) Competitive firms should get together to restrict output and drive up the price. (b) Firms would not advertise unless they expected advertising to increase sales. (c) A firm in a monopolistically competitive market faces a horizontal demand curve for its product.

MEDIUM

- 6 The table below shows the three-firm *CR* (concentration ratio) and the *NP* (market size divided by minimum efficient scale) for various industries in the UK, France and Germany in the mid-1970s. Refer to the table and answer the following questions:

Concentration and scale economies

Industry	UK		France		Germany	
	CR	NP	CR	NP	CR	NP
Refrigerators	65	1	100	2	72	3
Cigarettes	94	3	100	2	94	3
Refineries	79	8	60	7	47	9
Brewing	47	11	63	5	17	16
Fabrics	28	57	23	57	16	52
Shoes	17	165	13	128	20	197

- (a) True or false: In the fabric industry in Germany, the top three firms accounted for more than half of the market share.
- (b) True or false: In the shoe industry in the UK, France and Germany, economies of scale were relatively unimportant.
- (c) What can you conclude about the structure of the refrigerator and petroleum industries in the mid-1970s?
- (d) In the cigarette and petroleum industries, were there many firms that operated at minimum efficient scale? Explain your answer.

7 An industry faces the demand curve:

Q	1	2	3	4	5	6	7	8	9	10
P	10	9	8	7	6	5	4	3	2	1

- (a) Suppose it is a monopoly whose constant $MC = 3$: what price and output are chosen?
- (b) Now suppose there are two firms, each with $MC = AC = 3$: what price and output maximize joint profits if they collude?
- (c) Why might each firm be tempted to cheat if it can avoid retaliation by the other?
- 8 With the above industry demand curve, two firms, A and Z, begin with half the market each when charging the monopoly price. Marginal cost is constant at $MC = £3$.

Now, Z decides to cheat and believes A will stick to its old output level. (a) The following table shows the industry demand curve that Z thinks it faces. Compute the total revenue and marginal revenue and complete the table. (b) What price and output would Z then choose?

Q	1	2	3	4	5	6	7
P	8	7	6	5	4	3	2

- 9 Consider two firms, A and B. They have two possible strategies, pricing low or pricing high. The profits that each firm makes from those strategies are reported below:

		Firm B	
		P high	P low
Firm A	P high	3,3	1,6
	P low	6,1	5,5

What is meant by a Nash equilibrium? Find the Nash equilibrium for the strategies in the grid above.

- 10 Consider a market with two firms, 1 and 2, producing a homogeneous good. The market demand is $P = 100 - 3(Q_1 + Q_2)$, where Q_1 is the quantity produced by firm 1 and Q_2 is the quantity produced by firm 2. The total cost for firm 1 is $TC_1 = 40Q_1$, while the total cost for firm 2 is $TC_2 = 40Q_2$. Each firm behaves like a competitive firm.

- a) What is the equilibrium quantity in the market?
- b) Suppose both firms exhibit Cournot behaviour. Given that their reaction functions are $Q_1 = 20 - 2Q_2$ and $Q_2 = 20 - 2Q_1$, how would their output change compared to (a)?
- 11 Vehicle repairers sometimes suggest that mechanics should be licensed so that repairs are done only by qualified people. Some economists argue that customers can always ask whether a mechanic was trained at a reputable institution without needing to see any licence. (a) Evaluate the arguments for and against licensing car mechanics. (b) How would licensing affect the market for mechanics? (c) Are the arguments the same for licensing doctors?

HARD

- 12 **Essay question** ‘Globalization, by increasing the size of the market, reduces market power of individual firms and the need to address strategic interactions.’ ‘Globalization makes mergers more attractive and thus enhances worries about market power.’ Is either of these views correct? Or are both correct?

- |3 Two identical firms, 1 and 2, compete on quantities. The reaction function of firm 1 is $Q_1 = 15 - \frac{1}{2} Q_2$, while for firm 2 we have $Q_2 = 15 - \frac{1}{2} Q_1$. In the table below we have the total quantity produced in the market:

$Q_1 + Q_2$	2	6	10	14	18	22	26	30	34
-------------	---	---	----	----	----	----	----	----	----

Using the fact that both firms must produce the same quantity, plot the reaction functions of the two firms in a graph. How is the equilibrium quantity determined?

- |4 Consider a market with two firms, 1 and 2 producing a homogeneous good. The market demand is $P = 130 - 2(Q_1 - Q_2)$, where Q_1 is the quantity produced by firm 1 and Q_2 is the quantity produced by firm 2. The total cost for firm 1 is $TC_1 = 10Q_1$, while the total cost for firm 2 is $TC_2 = 10Q_2$. Each firm chooses the quantity to best maximize profits.
- (a) From the condition $MR_1 = MC_1$, find the reaction function of firm 1, and from $MR_2 = MC_2$, find the reaction function of firm 2.
 - (b) Find the equilibrium quantity produced by each firm by solving the system of the two reaction functions you found in (a). Sketch your solution graphically.
 - (c) Find the equilibrium price and then find the profit of each firm.

- 1 Another common measure for market concentration is the Herfindahl–Hirschman Index (HHI). For a given market this index is calculated as the sum of the square of the market share of each firm in the market. This index varies from 0 (no concentration) to 1 (the market is a monopoly).
- 2 Nash, who battled schizophrenia, won the Nobel Prize in Economics for his work on game theory. A film about his life, *A Beautiful Mind* (dir. Ron Howard), was released in 2001 and starred Russell Crowe.
- 3 The game, called the Prisoner's Dilemma, was first used to analyse the choice facing two people arrested and in different cells, each of whom could plead guilty or not guilty to the only crime that had been committed. Each prisoner would plead innocent if only she knew the other would plead guilty. For more information, go to www.mcgraw-hill.co.uk/textbooks/begg where there is a video fully explaining the Prisoner's Dilemma.

CHAPTER10

The labour market

Learning outcomes

By the end of this chapter, you should be able to:

- 1 analyse a firm's demand for inputs in the long run and short run
- 2 recognize marginal value product, marginal revenue product and marginal cost of a factor
- 3 define the industry demand for labour
- 4 analyse labour supply decisions
- 5 define economic rent
- 6 define labour market equilibrium and disequilibrium
- 7 understand how minimum wages affect employment

In winning a golf tournament, a top professional earns more in a weekend than a professor earns in a year. Students studying economics can expect higher career earnings than those of equally smart students studying philosophy. An unskilled worker in the EU earns more than an unskilled worker in India. Few market economies provide jobs for all their citizens wanting to work. How can we explain these aspects of the real world?

In each case the answer depends on the supply and demand for that type of labour. In this chapter and in the next one we analyse the markets for the factors of production – labour, capital and land. We discuss what determines the equilibrium prices and quantities of these inputs in different industries and in the whole economy. In this chapter we deal with the factor called ‘labour’. Chapter 11 applies the same principles to the markets for other production inputs.

We have already studied the market for goods. There is nothing intrinsically different about our approach to factor markets. You should be able to guess the structure of this chapter: demand, supply, equilibrium, problems of disequilibrium and adjustment.

Table 10.1 gives data on UK weekly earnings in 2012 and compares these with inflation-adjusted data for 2000. By 2012, workers in financial services earned £126 a week more than the national average. Workers in the construction sector saw their weekly real wage increase by 5 per cent during this period, an increase below the average increase for the whole economy (6.4 per cent). This is due to stagnation in the construction sector following the 2007 crisis.

Table 10.1 Weekly real earnings, UK (2012 £s)

	2000	2012
Whole economy	447	475
Financial services	573	621
Construction	527	554
Hotels and restaurants	305	333

Source: Annual Survey of Hours and Earnings, ONS.

Although the economics of factor markets still focus on supply and demand, there is something special about demand in factor markets. It is not a direct or final demand, but a **derived demand**. Firms demand inputs in order to produce output. For each firm, the decisions on how many inputs to demand and how much output to supply are inextricably linked.

The demand for inputs is a **derived demand**, reflecting demand for the firm's output.

On the supply side, we distinguish between the supply of factors to the economy and to an individual firm or industry. A firm can gain inputs by attracting them away from other firms. However, the economy as a whole may be able to expand particular inputs only slowly. It takes time to build factories or train skilled workers.

In the short run, the supply of pilots to the economy may be fixed. Any rise in the total demand for pilots raises their equilibrium wage. In the longer run, high wages for pilots act as a signal for school-leavers to

abandon plans to become train drivers and go to flying school instead. Thus, we need to distinguish between labour supply in the short run and long run.

Combining demand and supply leads to equilibrium prices and quantities in the labour market. How quickly does the labour market return to equilibrium? Whereas some output markets may return to equilibrium relatively quickly, labour market adjustment is often more sluggish. We examine reasons why the labour market may be slow to adjust.

10.1 The firm's demand for factors in the long run

In the long run, all inputs can be adjusted. Chapter 7 studied a firm's long-run costs. Chapters 8 and Chapter 9 considered various descriptions of the demand curve facing a firm and showed how a firm would choose output supplied to maximize profits. Although these are part of the same decision, we now focus not on the firm's supply of output but on its corresponding demand for inputs.

The firm thinks about the least-cost way of making each possible output and then selects the output that maximizes profit. In producing any particular output by the cheapest available technique, a rise in the price of labour relative to capital makes the firm switch to a more capital-intensive technique. Conversely, if capital becomes relatively more expensive, the least-cost technique for a given output is now more labour-intensive. The firm substitutes away from the factor of production that has become relatively more expensive.

This principle helps explain cross-country differences in capital-labour ratios in the same industry. European farmers face high wages relative to the rental of a combine harvester. Mechanized farming economizes on expensive workers. Indian farmers, facing cheap and abundant labour but scarce and expensive capital, use labour-intensive techniques. Workers with scythes and shovels do the jobs done by combine harvesters and bulldozers in the UK.

A higher wage makes the firm substitute capital for labour in making a given output. But it also raises the total cost of producing any output. Firms still use *some* labour, for which they now pay more than before.

With higher marginal costs, but unchanged demand and marginal revenue curves, the firm chooses to make less output.

Thus a rise in the price of one factor changes not only factor intensity at a given output but also the profit-maximizing level of output. Studying consumer decisions in Chapter 5, we saw that a change in the price of a good has both a substitution effect and an income effect. The substitution effect reflects the change in relative prices of different goods and the income effect reflects changes in real income as a result of the price change. The demand for production inputs works in exactly the same way.

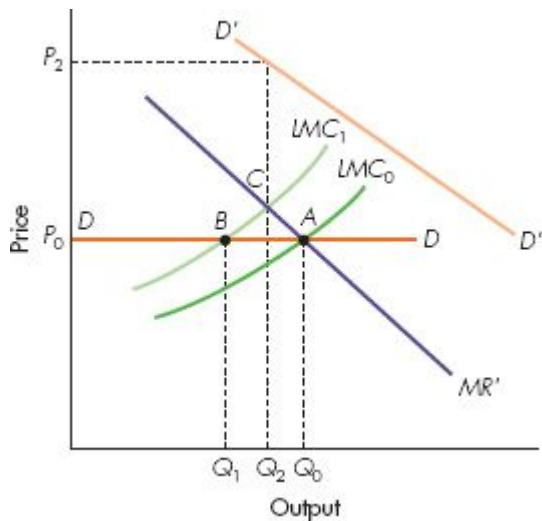
There is a pure substitution effect at a given level of output. A higher relative price of labour compared with capital leads firms to substitute capital for labour. But there is also an output effect; that is, the analogue of the income effect in consumer demand theory. By raising the marginal cost of producing output, a rise in the price of labour leads to a lower output.

In the long run, a rise in the wage *will* reduce the quantity of labour demanded. The substitution effect leads to less demand for labour and each output, and the output effect reduces the demand for all inputs.

A rise in the wage also affects the long-run demand for capital and other inputs. At any particular output, the firm substitutes capital for labour. However, with lower output it needs less capital input. The overall effect could go either way. The easier it is to substitute capital for labour, the more likely is the substitution effect to dominate. Firms will substitute a lot of capital for labour. The quantity of capital demanded will rise.

The demand for factors of production is a derived demand. It depends on demand for the firm's output. The output demand curve affects the output effect on the demand for inputs when an input price changes.

In Figure 10.1 at the original wage, the long-run marginal cost curve LMC of output is LMC_0 . A rise in the wage shifts this up to LMC_1 . The original profit-maximizing point is A . If the firm faces a horizontal demand curve DD , output falls from Q_0 to Q_1 . With the less elastic demand curve $D'D'$, the firm still begins at A where LMC_0 equals MR , the marginal revenue curve corresponding to $D'D'$. Now the shift to LMC_1 leads to a much smaller fall in output. The new output is Q_2 and the firm is at C .



A wage increase will have a substitution effect leading firms to substitute relatively more capital-intensive techniques. Nevertheless, total costs and marginal costs of producing output will be greater than before. Facing the horizontal demand curve DD, a shift from LMC_0 to LMC_1 will lead the firm to move from A to B and output will fall from Q_0 to Q_1 . This tends to reduce the demand for all factors of production. Facing the demand curve $D'D'$ and corresponding marginal revenue curve MR' , the upward shift from LMC_0 to LMC_1 leads the firm to move from A to C at which marginal cost and marginal revenue are again equal. The output effect reduces output only from Q_0 to Q_2 .

Figure 10.1 The output effect of a wage increase

The more elastic the demand curve for the firm's output, the more a given rise in the price of an input, and a given shift in the LMC curve for output leads to a big fall in output. The larger the output effect, the greater the fall in the quantity of all factors demanded.

The Appendix at the end of this chapter shows that we can also analyse factor demands using techniques such as the indifference curves and budget lines used to study household demands for goods in Chapter 5.

10.2

The firm's demand for labour in the short run

In the short run, the firm has some fixed factors of production. We now consider the firm's short-run demand for labour when its capital input is fixed.

Table 10.2 shows a firm's variable labour input and corresponding output, holding capital input fixed. Column (3) shows the **marginal product of labour (MPL)**. This marginal product rises as the first workers are added. It is hard for the first and second worker to carry all the tools. After the third worker has been added, the *diminishing marginal productivity* of labour sets in. With existing machines fully utilized, there is less and less for each new worker to do.

The **marginal product of labour (MPL)** is the extra total output when an extra worker is added, with other input quantities unaltered.

Table 10.2 Short-run output supply and labour demand

(1) Workers	(2) Output	(3) <i>MPL</i>	(4) <i>MVPL (£)</i>	(5) Extra profit (£)
1	0.8	0.8	400	100
2	1.8	1.0	500	200
3	3.1	1.3	650	350
4	4.3	1.2	600	300
5	5.4	1.1	550	250
6	6.3	0.9	450	150
7	7.0	0.7	350	50
8	7.5	0.5	250	250

As in our discussion of output, we use the *marginal principle*. Does the cost of a new worker exceed the benefit of a new worker? Table 10.2 shows a competitive firm hiring workers at a wage of £300 and selling output at a price of £500. Column (4) shows the extra revenue from taking on another worker.

Since the firm is perfectly competitive, the marginal value product of another worker is the marginal product in physical goods multiplied by the (constant) price for which the extra goods are sold. From this extra revenue from the extra worker, the firm subtracts the extra wage cost. The last column of Table 10.2 shows the extra profit from an extra worker.

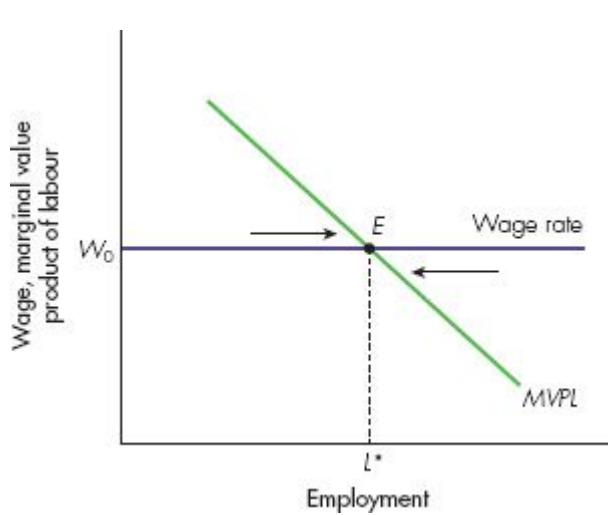
The firm hires more workers if the **marginal value product** of another worker exceeds the wage cost. It is profitable to hire 7 workers. The

seventh worker has a marginal value product of £350, just above the cost of £300 for this extra worker. An eighth worker's marginal value product is only £250; that is, below the £300 another worker costs. The firm hires 7 workers.

The **marginal value product of labour** is the extra revenue from selling the output made by an extra worker.

In so doing, the firm chooses both labour input and goods output: the highlighted row shows that 7 workers make 7 units of output. The firm gets the same answer, namely maximum profit, whether it compares the marginal revenue from another output unit with the marginal cost of making that output unit, or compares the marginal revenue from hiring another unit of the variable factor with the marginal cost of hiring that variable factor.

The firm's employment rule is thus: expand (contract) employment if the marginal value product of labour is greater than (less than) the wage of an extra worker. If labour can be smoothly adjusted, for example if labour input is measured in [hours] 3 [workers], the firm's demand for labour must satisfy the condition



The firm sells output for a given price and hires labour at the given wage W^* . Diminishing marginal productivity makes the MVPL schedule slope down. Below L^* extra employment adds more to revenue than labour costs. Above L^* extra employment adds more to costs than to revenue. L^* is the profit-maximizing employment level where the wage rate equals the MVPL

Figure 10.2 The firm's choice of employment

$$\text{Wage} = \text{marginal value product of labour (}MVPL\text{)} \quad (1)$$

Figure 10.2 illustrates this principle. If we assume diminishing marginal productivity at all employment levels, the marginal value product of labour ($MVPL$) slopes down. A competitive firm can hire labour at the constant wage rate W_0 . It is a price-taker in the labour market. Below L^* , profits are increased by raising employment, since $MVPL$ exceeds the wage rate or marginal cost of hiring extra labour. Above L^* , it is profitable to shrink employment, since the wage exceeds the $MVPL$. Thus L^* is the profit-maximizing level of employment.

Changes in the firm's demand for labour

Consider a rise in the wage W_0 faced by a competitive firm. Using Figure 10.1 or 10.2, the firm hires fewer workers than before. The marginal cost of labour has risen. Diminishing labour productivity makes the $MVPL$ schedule slope down. Hence lower employment is needed to raise the marginal value product of labour in line with its higher marginal cost.

Suppose that a competitive firm faces a higher output price. The MPL remains unaltered in physical goods, but this output now earns more money. The $MVPL$ schedule shifts up at each level of employment. Hence in Figure 10.1 or 10.2 the horizontal line through the wage W_0 crosses the new $MVPL$ schedule at a higher employment level. With the marginal cost of labour unaltered and the marginal revenue from labour increased, output and employment expand until diminishing marginal productivity drives $MVPL$ back down to the wage W_0 .

Finally, suppose the firm had begun with a higher capital stock. Each worker has more machinery with which to work and makes more output. Although wages and prices are unchanged, there is a rise in MPL in physical goods at each employment level. The $MVPL$ schedule shifts up, since $MVPL$ equals MPL multiplied by output price. As with a higher output price, this upward shift in the $MVPL$ schedule leads the firm to expand employment and output.

For a competitive firm there is a neat way to combine our first two results. Noting that $MVPL$ equals the output price P multiplied by MPL ,

the extra physical product of another worker, the firm's profit-maximizing condition is wage $W = P \times MPL$. Dividing both sides of this equation by P gives:

$$W/P = MPL \quad (2)$$

A profit-maximizing competitive firm demands labour up to the point at which the marginal physical product of labour equals its *real* wage, the nominal wage divided by the output price.

The position of the *MPL* schedule depends on technology and the existing capital stock. Since these are fixed in the short run, we can alter *MPL* only by moving along the schedule. Diminishing returns imply that, with more workers, the marginal physical product of the last worker is lower. From the particular level of the marginal physical product of labour, we can deduce how many workers are being employed.

Equation (2) tells us that if nominal wages and output prices both double, real wages and employment are unaffected. But changes in either the nominal wage or the output price, if not matched by a change in the other, alter employment by affecting the real wage. Lower real wages move the firm down its *MPL* schedule, taking on more workers until the marginal physical product of labour equals the real wage.

Having studied the firm's demand for labour in the short run, we now turn to the demand of the industry as a whole. Although each competitive firm regards itself as a price-taker in both its output and input markets, an expansion by the whole industry will change output prices and wages. In moving from the firm's demand curve to the industry demand curve for labour, we take account of these effects.

MATHS 10.1

THE DEMAND FOR INPUTS AND PROFIT MAXIMIZATION

Here we see how to derive the demand for inputs from profit maximization in competitive markets. We consider both labour and capital, even though capital will be discussed in more detail in Chapter 11.

Consider a perfectly competitive firm that faces competitive markets for capital and labour. The firm is producing output

according to the following production function:

$$Q = f(K, L)$$

where K is the amount of capital and L is the amount of labour.

The prices of inputs are r for capital (the rental rate) and w for labour (the wage) and they are taken as given by the firm. The profit function of the firm is:

$$\pi = pf(K, L) - wL - rK$$

where p is the price (taken as given) at which the firm sells its output Q . The firm chooses the amounts of capital and labour to employ in order to maximize profits.

We know that this implies that the firm chooses K and L , such that the marginal value product of the two inputs equals the prices of the two inputs:

$$P \frac{\partial f(K, L)}{\partial K} = r \quad P \frac{\partial f(K, L)}{\partial L} = w$$

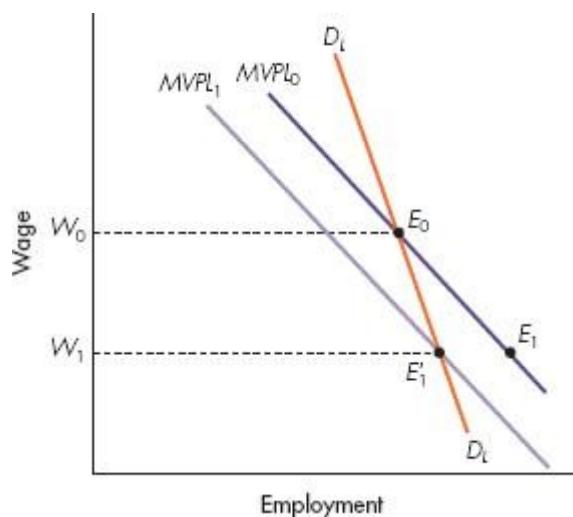
The term $(\partial f(K, L))/\partial L$ is the *partial derivative* of the production function with respect to labour. It tells us how the quantity produced (remember that $Q = f(K, L)$) changes if we change the level of labour by a small amount, *keeping constant* the level of capital. If labour increases by 1 unit, for a given amount of capital, the extra output we obtain is $(\partial f(K, L))/\partial L$. We can sell this extra output at a price p . Therefore the term $p((\partial f(K, L))/\partial L)$ is the marginal value product of labour. A similar interpretation is given for the term $p((\partial f(K, L))/\partial K)$, which represents the marginal value product of capital.

10.3 The industry demand curve for labour

For a given price P_0 and wage W_0 , each firm in a competitive industry chooses employment to equate the wage and the *MVPL*. Figure 10.3 horizontally adds the marginal value product of labour curves for each

firm to obtain the $MVPL_0$ schedule for the industry. At the wage W_0 and the price P_0 , the industry is at E_0 . This is a point on the industry demand curve for labour.

However, $MVPL_0$ is *not* the industry demand curve for labour. It is drawn for a particular output price P_0 . Suppose the wage is cut from W_0 to W_1 . At the output price P_0 , each firm wants to move down its $MVPL$ schedule and the industry expands output and employs labour to point E_1 in Figure 10.3. In terms of the supply and demand for output, the cut in wages has shifted the industry supply curve to the right.



$MVPL_0$ is the horizontal sum of each firm's $MVPL$ schedule at the price P_0 . Each firm and the industry as a whole sets $MVPL$ equal to W_0 . Hence E_0 is a point on the industry demand curve for labour. A lower wage W_1 leads each firm and the industry as a whole to move down their $MVPL$ schedules to a point E_1 . Extra employment and output by the whole industry (a shift to the right in the industry supply curve of goods) leads to excess goods supply at the original price P_0 . To clear the output market the price must fall, and this shifts to the left each firm's $MVPL$ schedule. The new industry schedule is $MVPL_1$ and the chosen point is E'_1 . Joining all the points such as E_0 and E'_1 , we obtain the industry demand curve $D_L D_L$.

Figure 10.3 The industry demand for labour

At the given price P_0 , there is now an excess supply of goods. This bids down the price for the industry's product to a lower price P_1 . The lower price shifts each firm's $MVPL$ schedule to the left. $MVPL_1$ is thus the

new $MVPL$ schedule for the industry at the new price P_1 . The industry chooses the point E_1' at the new wage W_1 .

Connecting points such as E_0 and E_1' , we get the *industry demand for labour schedule* $D_L D_L$ in Figure 10.3. Each firm constructs its $MVPL$ schedule as if it were a price-taker but the industry demand curve has a steeper slope, since a lower wage shifts the industry output supply curve to the right and reduces the equilibrium price.

The slope of the $MVPL$ schedule reflects the production technology. The more MPL diminishes as labour input rises, the steeper is the $MVPL$ schedule of the firm and of the industry. The slope of the industry demand curve for labour also depends on the elasticity of the market demand curve for the industry's product. The more inelastic is output demand, the more a wage cut – by raising the supply of output – bids down the output price and shifts $MVPL$ schedules to the left , and the steeper is the industry demand curve $D_L D_L$ for labour.

The demand for factors of production is a *derived* demand. Firms want factors only because they see a demand for their output that it is profitable to supply. The elasticity of input demand reflects the elasticity of output demand.

10.4 The supply of labour

We now discuss the supply of labour, for the individual, the industry and the whole of the economy. We can then combine labour demand and labour supply to determine the equilibrium level of wages and employment.

Individual labour supply: hours of work

We analyse labour supply in two stages: how many hours people work once in the **labour force** and whether people join the labour force at all.

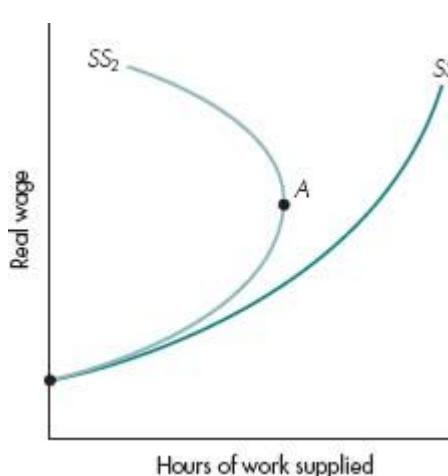
The **labour force** is all individuals in work or looking for work.

Once in the labour force, how many hours will a person wish to work? This depends on the *real wage*, W/P , the nominal wage divided by the price of goods, which shows the quantity of goods that labour effort will purchase. It is the real wage that affects labour supply decisions.¹

Figure 10.4 shows two possible labour supply curves, relating hours of work supplied to the real wage. The curve SS_1 slopes up. As real wages increase, people want to work more. The labour supply curve SS_2 is *backward-bending*. Beyond A , further real wage rises make people want to work fewer hours.

The alternative to working another hour is staying at home and having fun. Each of us has 24 hours a day to divide between work and leisure. More leisure is nice but by working longer we can get more real income with which to buy goods. How should an individual trade off leisure against consumer goods in deciding how much to work?

This is an application of the model of consumer choice in Chapter 5. The choice is now between goods as a whole and leisure. An individual will want to work until the marginal utility derived from the goods that an extra hour of work will provide is just equal to the marginal utility from the last hour of leisure.



The labour supply curve SS_1 slopes up and more hours of work are supplied as the real wage increases. But the labour supply curve might bend back. Along SS_2 higher real wages reduce labour supply once we reach the point A.

Figure 10.4 Individual labour supply

A higher real wage increases the quantity of goods an extra hour of work will purchase. This makes working more attractive than before and tends to increase the supply of hours worked. But there is a second effect. Suppose you work to get a target bundle of goods. You work to get enough to be able to eat, pay the rent, run a car and have a holiday. With a higher real wage you need to work fewer hours to earn the same target bundle of goods.

These two effects are precisely the *substitution and income effects* introduced in the consumer choice model of Chapter 5. An increase in the real wage increases the relative return on working. It leads to a substitution effect that makes people want to work more. But a higher real wage also tends to raise people's real income. This has a pure income effect. Since leisure is probably a luxury good, the quantity of leisure demanded increases sharply when real incomes increase. This income effect tends to make people work less. The overall effect of a real wage rise, and the shape of the supply curve for hours worked, depends on which effect is larger.

To decide whether or not the substitution effect will dominate the income effect, we must look at actual data on what people do. The empirical evidence for the UK, the US and most other Western economies is as follows. For adult men, the substitution effect and the income effect almost exactly cancel out. A change in the real wage has almost no effect on the quantity of hours supplied. The supply curve of hours worked is almost vertical.²

For women, the substitution effect just about dominates the income effect. Their supply curve for hours worked slopes upward. Higher real wages make women work longer hours.

Workers care about take-home pay after deductions of income tax. Lower income tax rates raise after-tax real wages. The empirical evidence on labour supply implies that lower income tax rates should not be expected to lead to a dramatic increase in the supply of hours worked.

MATHS 10.2

INDIVIDUAL LABOUR SUPPLY AND INDIFFERENCE CURVES

Here we see how to derive the labour supply of an individual worker using the tools developed in Chapter 5. In a given day an individual can spend h hours on work and l hours on leisure. Leisure is any time in a day spent not working. Time is a scarce resource since there are only 24 hours a day. The time constraint faced by our individual is:

$$24 = l + h \quad (1)$$

We assume that our individual derives utility from consumption of various goods and from leisure. Working creates a *disutility*. He prefers leisure from which he derives utility. We assume that without working our individual's income is zero (the case in which there is a positive non-labour income does not change the main results; an example is provided in Figure 10.5). In order to consume, our individual must work.

The hourly *real* wage our individual earns is given by w . He takes this wage as given. Labour income is $I = wh$, which is the hourly wage multiplied by the number of hours worked. C is the amount of consumption. He consumes all his labour income and therefore $C = wh$.

$$h = \frac{C}{w}$$

The latter equation can be written as $C = wh$. We can substitute this equation into (1) and re-arrange it to obtain the individual's budget line:

$$C = 24w - wl \quad (2)$$

Equation (2) tells us the feasible combination of consumption and leisure our individual can achieve given the wage w . Notice that the budget line is negatively sloped (the slope is $-w$), denoting the existence of a trade-off between consumption and leisure. To increase consumption, our individual must spend more hours working and decrease his leisure time. Notice that the wage denotes the *opportunity cost* of leisure. To increase leisure by one extra hour, our individual must give up an hour of work that pays w .

The individual's preferences regarding consumption and leisure are summarized by standard indifference curves. The problem he faces is: given his preferences and his budget line, how many hours

should he choose for leisure and how much should he consume (that is, how many hours should he work)? The solution lies where the budget line is tangent to the highest possible indifference curve. This is shown as point A in the figure below, where leisure is on the horizontal axis and consumption is on the vertical axis.

At point A , the marginal rate of substitution between consumption and leisure (that is, the slope of the indifference curve) is equal to the hourly real wage (that is, the slope of the budget line). Our individual will choose l^* hours of leisure and therefore will choose to work $h^* = 24 - l^*$ hours and will consume an amount $C^* = wh^*$.

Now suppose that the hourly real wage increases to $w_1 > w$. The increase in the wage has the effect of rotating the budget line outwards around the point on the horizontal axis where $l=24$. The new budget line is given by the line connecting the points $l = 24$ and $C = 24w_1$. The optimal choice of our individual is now given by point B , where the new budget line is tangent to the indifference curve I_2 .

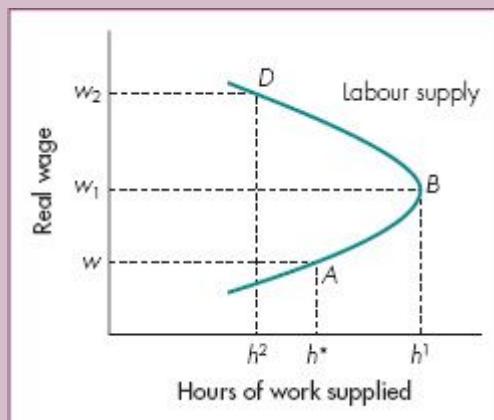
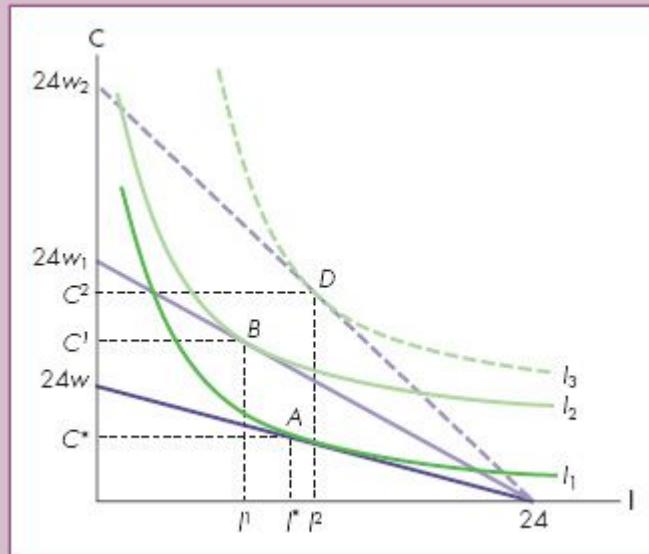
At point B , the individual chooses l^1 hours of leisure, meaning he chooses to work $h^1 = 24 - l^1$ hours and consume $C^1 = w_1h^1$. As we know, the increase in the wage has two effects, a substitution effect and an income effect.

Comparing point A with point B , we can see that after the wage increase our individual increases the number of working hours ($h^1 > h^*$), meaning that the substitution effect dominates. Moreover, by working more his consumption increases ($C^1 > C^*$).

Suppose the real wage increases further, to w_2 . The new optimal choice is at point D . Our individual now chooses l^2 hours of leisure; he works $h^2 = 24 - l^2$ hours and consumes $C^2 = w_2h^2$. Compared to point B , he reduces the number of working hours ($h^2 < h^1$) and increases consumption. In moving from point B to point D , the income effect dominates. The wage he gets is so high that he can still achieve high income and thus also high consumption by working less than before.

The labour supply of our individual tells us how many hours he would like to work at different wages. Given the information

derived above, we can plot our individual's labour supply as shown in the figure below.



The labour supply of our individual is backward-bending. Starting at a lower wage, an increase in the wage will result in an increase in hours worked. The substitution effect dominates. As the wage increases further, labour supply decreases as the income effect starts to dominate.

Whether the labour supply is backward-bending depends on an individual's preferences regarding leisure and consumption. Using the same analysis we could have specified the indifference curves in such a way that the resulting labour supply would have been upward-sloping, like SS_1 in Figure 10.4.

Individual labour supply: participation rates

The effect of real wages on the supply of hours is smaller than often supposed. The more important effect of real wages on labour supply is on the incentive to join the labour force.

Table 10.3 gives data on **participation rates** for different countries in 1994 and 2011. Most men of working age are in jobs or are seeking employment, but this percentage is gradually falling in some countries like the UK and the US, while in others it is slightly increasing. On the other hand, there has been a rise in labour force participation by women in the last 17 years. Can our model of choice explain these trends?

The **participation rate** is the fraction of the working age population who join the labour force.

Table 10.3 Participation rates (%)

	1994		2011	
	Men	Women	Men	Women
UK	85.1	67.1	82.7	70.4
France	74.1	59.3	74.7	66.1
Germany	79.8	60.9	82.6	71.8
US	84.3	69.4	78.9	67.8
EU15	78.4	56.5	79.2	65.5

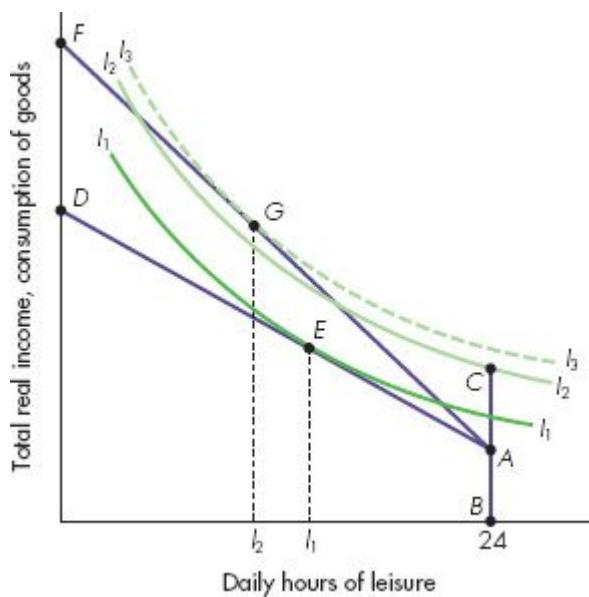
Note: EU15 refers to the average for the 15 countries that were in the European Union on 1 January 1995.

Source: Based on data from Labour Force Survey, © OECD, 2012, www.oecd.org/employment/outlook, accessed on 18/06/2013.

We now develop a model in which labour force participation is higher (a) the more people's tastes favour the benefits of working (goods or job status) relative to the benefits of leisure, (b) the lower their income from non-work sources, (c) the lower the fixed costs of working, and (d) the higher the real wage rate.

Figure 10.5 plots leisure on the horizontal axis. Twenty-four hours is the maximum amount of leisure that can be enjoyed in a day. The vertical axis plots total real income from work and other sources. This shows the ability to buy goods and services. We begin with the budget constraint.

Suppose the individual has a non-labour income given by the vertical distance BC . This may be income earned by a spouse, income from rent or dividends, or welfare payments received from the government.



With a non-labour income BC the individual can do no work and consume at C on the indifference curve I_2I_2 . Any work incurs the fixed cost AC . At a low hourly wage rate the total budget line is CAD and the best point attainable by working is E . Thus this lies on the indifference curve I_1I_1 and the individual is better off at C where no work is done. At a higher hourly wage rate the new budget line is CAF . By working $(24 - l_2)$ hours the individual can reach G on the indifference curve I_3I_3 , which is better than being at C . The higher the real hourly wage rate, the more likely is the individual to participate in the labour force.

Figure 10.5 Labour force participation

Someone not working at all can have 24 hours of leisure a day plus a daily income BC . He can consume at point C . Now suppose he works. There may be fixed costs in working. Unemployment benefit from the government may be lost immediately, the right clothes or uniform must be purchased, travel expenses must be incurred to get to the place of work and childcare must be found for the children. These costs are independent of the number of hours worked, provided any work is done. They are a fixed cost of working.

Figure 10.5 shows these costs as the vertical distance AC . Instead of being able to consume at C , the net non-labour income BC is reduced to BA after these fixed costs of working are incurred. Having decided to

work, he can then move along the budget line AD , sacrificing leisure to gain wage income. The higher the real wage, the steeper the budget line AD .

Fixed costs of working lead to a kinked budget line CAD . Working a few hours reduces total real income. The small wage income does not cover the fixed costs of working. The lower the real wage rate, the flatter is the AD line and the more hours he has to work merely to recoup the fixed costs. This is sometimes called the *unemployment trap*. Unskilled workers face such a low wage that they actually lose out by working.

To complete the model of consumer choice, we superimpose an indifference map on the kinked budget line CAD . Individuals like both leisure and goods. Each indifference curve has the usual slope and curvature. A higher indifference curve means the individual is better off. We can now analyse the participation decision and establish the four effects we cited above.

The indifference curve I_2I_2 shows how well off the individual is as a result of not participating. He can start to consume at C . Given the budget line CAD , the best he can do by working is to work ($24 - 2l_1$) hours, consume l_1 hours of leisure and choose point E , reaching the indifference curve I_1I_1 . But he can reach the higher indifference curve I_2I_2 by not working. He chooses not to work.

Now suppose the real wage rises. Each hour of leisure could now earn a higher real wage. AD rotates to AF and the complete budget line is now CAF . By choosing point G , he can reach the indifference curve I_3I_3 and is better off than at C . Hence higher real wages raise the number of people wishing to join the labour force.

This analysis implies that, if the real wage is sufficiently low, our individual decides not to work. As the real wage becomes sufficiently high he is going to participate in the labour market. This means that there must be a real wage at which our individual is indifferent between working and not working. That real wage is called the *reservation wage*. The reservation wage is the lowest wage a worker is willing to accept to work in a given occupation. To participate in the labour market, he must be offered at least his reservation wage.

A reduction in AC , the fixed cost of working, also raises participation. Point C is fixed but point A shifts up. There is a parallel upward shift in

the sloping part of the budget line such as AD or AF . It is more likely that the highest indifference curve attainable by working will lie above the zero-work indifference curve I_2I_2 .

Although not shown in Figure 10.5, lower non-labour income BC also raises labour force participation. This raises the opportunity cost of not working, making working more attractive. Changes in non-labour income have no effect on the relative return of an hour's work and an hour's leisure. There is no substitution effect but there is an income effect. Lower non-labour income reduces the quantity demanded of all normal goods, including leisure. People are more likely to work.

Finally, consider a change in tastes. People decide leisure is less important and work more important. Each indifference curve in Figure 10.5 is flatter: people are prepared to sacrifice more leisure for the direct and indirect benefits of extra work. Consider again the budget line CAD . The flatter are the indifference curves, the more likely it is that the indifference curve through C will cross the portion of the budget line AD on which work is done. But if it crosses AD there must be another point on AD yielding even higher utility. In Figure 10.5 it is possible to attain a higher indifference curve by choosing point G on AF . Exactly the same argument applies if the flatter indifference curve through C crosses the line AD .

CASE 10.1

THE PUZZLE OF LOW FEMALE PARTICIPATION IN THE LABOUR FORCE IN TURKEY

Turkey is a large and growing economy. In the 1980s the participation rate of women in the Turkish labour market was the same as in Austria, the Netherlands and Switzerland. However, over time, the female participation rate showed a negative trend, moving from 34.3 per cent in 1988 to 22 per cent in 2008.

During the same period Turkey experienced structural changes that should have facilitated the entry of women in the labour market. Women are becoming better educated and are getting married at a later age. The fertility rate (that is, the average number of children per woman) has declined and the social attitude towards working

women has changed in comparison to the past. Thus the decline in participation rate seems puzzling.

According to a recent report, urbanization and the decline in agricultural employment are the main factors explaining the decline in female participation in the labour market in Turkey. Women in rural areas participate in the labour market mainly by engaging in agricultural activities. However, agricultural employment in Turkish rural areas has decreased. Young men are becoming more educated and thus move away from agricultural employment to better paid jobs. Shifting family activity away from agricultural employment causes a withdrawal of wives from the labour force.

Moreover, over the last 30 years one-third of Turkish women have been internal migrants, moving from rural areas to urban areas. This means that women migrate from a high-participation rural environment to a low-participation urban environment (where many of them do not participate in the labour market and stay at home). More educated women benefit from urbanization and get better jobs. However, poorly educated women are disadvantaged by the move. Indeed the low female participation rate is mainly created by poorly educated women in urban areas. This situation may create what is called the ‘under-participation trap’. Women with low levels of education in urban areas are more likely to work in the informal sector at a very low wage. This decreases the participation of poorly educated women in the labour market. The prospect of low wages may cause families to under-invest in the education of girls, creating a vicious cycle.

Source: ‘Female labour force participation in Turkey: trends, determinants and policy framework’ World Bank, Report No 48508 – TR.

Thus labour force participation rises with (a) a higher real hourly wage rate, (b) lower fixed costs of working, (c) lower income from non-labour sources, and (d) changes in tastes in favour of more work and less leisure. Is this why participation in the labour market by married women increased?

First, there was a change in social attitudes to work, especially to work by married women. Indifference curves became flatter. Second, pressure for equal opportunities for women raised women’s real wages. The budget line for working women rotated from AD to AF in Figure 10.5.

Finally, the fixed costs of working fell. Labour-saving devices for housework, a second family car and many other changes, not least in the attitude of husbands, reduced the cost of work, especially for married women.

We have reached two conclusions. First, a higher real wage rate raises total labour supply but perhaps by less than is commonly thought. Second, this operates more by sucking people into the labour force than by raising the supply hours of those already in the labour force. This analysis relates best to the supply of unskilled workers.

The supply of labour to an industry

Now we discuss an individual industry (or market). Suppose it is small relative to the economy and wishes to employ workers with common skills. It has to pay the going rate for the job. Jobs in different industries have different non-monetary characteristics, such as risk, comfort, opportunity for promotion, and so on. The going rate must be adjusted industry by industry to allow for the *equilibrium wage differential* that offsets these non-monetary characteristics and makes workers indifferent to where they work. Dangerous, nasty industries have to pay more than safe, pleasant industries if they are to attract workers.

Adjusted in this way, this determines the wage at which a small industry can hire as many workers as it wants from the economy-wide labour pool. At this wage, the industry faces a horizontal labour supply curve.

Many industries are not this small relative to all the skills they wish to employ. The steel industry is a big user of welders, the freight industry a big user of lorry drivers. When an industry is a significant user of a particular skill, higher employment in the industry bids up the wages of that particular skill in the whole economy. In the short run, the industry's labour supply curve slopes upwards.³

In the long run, the industry's labour supply curve may be flatter. When short-run expansion bids up the wages of computer programmers, more school-leavers train in this skill. In the long run, the economy-wide supply rises and the wages of these workers fall back a bit. An individual industry does not have to offer such a high wage in the long run to increase the supply of that type of labour to the industry.

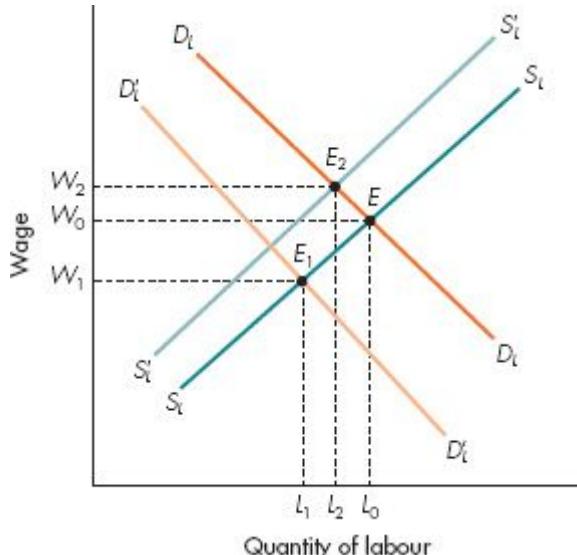
In the short run, the supply of a given skill may be nearly fixed. To get a larger share of the total pool, an individual industry has to offer higher

relative wages than other industries to bid workers away from them.

10.5

Industry labour market equilibrium

Figure 10.6 shows equilibrium in the labour market for an industry. Its labour demand curve $D_L D_L$ slopes down and crosses the upward-sloping labour supply curve $S_L S_L$ at the equilibrium point E . Employment is L_0 and the wage W_0 . We do not distinguish between long-run and short-run supply curves, though this is easily done.



The industry labour market supply curve $S_L S_L$ slopes up. Higher wages are needed to attract workers into the industry. For a given output demand curve, the industry's labour demand curve $D_L D_L$ slopes down because of diminishing marginal labour productivity and because higher industry output bids down its output price. A leftward shift in the output demand curve thus shifts the derived demand for labour from $D_L D_L$ to $D'_L D'_L$ and moves labour market equilibrium from E to E_1 . An increase in wages elsewhere in the economy shifts the industry's labour supply curve from $S_L S_L$ to $S'_L S'_L$ and shifts equilibrium from E to E_2 .

Figure 10.6 Equilibrium in an industry labour market

We draw the industry labour demand curve $D_L D_L$ for a given output demand curve. A recession in the building industry would shift the demand curve for cement to the left. The equilibrium price of cement

falls. This shift to the left the marginal value product of labour curve $MVPL$ for each cement manufacturer. Hence D_1D_1 shifts to $D'_1D'_1$ for the cement industry. At the new equilibrium E_1 , wages and employment are lower in the industry.

Conversely, suppose there is a spurt of investment in new machinery in every industry except cement. With more capital to work with, labour is more productive in other industries. Setting wages equal to the $MVPL$, these industries pay higher wages. This shifts up the supply curve of labour to the cement industry to $S'_LS'_L$. At each wage in the cement industry, the industry attracts fewer workers from the labour pool.

The new equilibrium for cement workers is at E_2 . Employment falls from L_0 to L_2 . Since the remaining workers have more capital to work with, they have a higher marginal product. In addition, the contraction in cement output shifts the output supply curve to the left and bids up the cement price. Together, these effects move the industry up its demand curve D_LD_L and allow it to pay a higher wage rate to its remaining workers.

Thus wage increases in one industry spill over into other industries. The crucial link between industries is *labour mobility*. It is because cement workers are lured away from the industry by wage rises elsewhere that the cement industry's labour supply curve shifts to the left in Figure 10.6. The degree of labour mobility between industries affects not only how much an industry's labour supply curve shifts when conditions change elsewhere, but also the slope of the industry's labour supply curve. Consider two extreme cases.

Suppose first that workers can move effortlessly between similar jobs in different industries. If each industry is small relative to the economy, it will face a completely elastic (horizontal) labour supply curve at the going wage rate (adjusted for non-monetary advantages). When all other industries pay higher wages, the horizontal supply curve of labour to the cement industry shifts up by the full amount of the wage increase elsewhere. Unless the cement industry matches the going rate, it loses all its workers.

At the opposite extreme, consider the market for concert pianists. Suppose they can do no other job. The supply curve of concert pianists is vertical. If all other industries pay higher wages, this has no effect on the

market for concert pianists. There is no possible entry into or exit from the occupation of concert pianists.

The general case of Figure 10.6 is between these extremes. With limited mobility between industries, the cement industry can attract more workers by offering higher wages. But its labour supply curve shifts when wages change elsewhere.

CASE 10.2

DOES IMMIGRATION HURT NATIVE WORKERS?

In many countries there is rising concern about the possible negative effects that immigration may have on wages and employment outcomes for native workers.

If we think about immigration as an increase in labour supply, for a given labour demand, we should expect a decrease in the equilibrium wage. If the wage decreases, some of the native workers may reduce their labour supply and this may reduce their employment level.

In the UK in the last decade immigration has increased sharply. The net inflow of immigrants increased from 50 000 individuals in 1995 to 220000 in 2005.

Recent research has tried to identify the possible effects on the labour market that immigration may have on native residents. Research by Dustmann et al. found that immigration has very little effect on the labour market outcomes of native workers. In particular, an increase in immigration seems not to have any significant effect on the native employment rate. In terms of wages, an increase in immigration amounting to 1 per cent of the native population would lead to just under a 2 per cent increase in average native wages. Those results confirmed some of the results found for the US economy. However, those results are related to the overall labour market in the UK. If we look in more detail at different occupations, the effect of immigration on native wages may instead be negative, at least for some of them. This is the finding of research by Nickell and Saleheen (2008). They found that, for jobs such as managers, skilled production workers (engineers, IT

technicians, and so on) and semi/unskilled service workers (cleaners, labourers, and so on), an increase in immigration has a small, but negative, effect on average wages of native workers in those occupations. Therefore, whilst overall the impact of immigration seems negligible, for some specific occupations immigration may have a negative effect on native workers.

Sources: C. Dustmann, F. Fabbri and I. Preston, ‘The impact of immigration on the British labour market’, *Economic Journal* 115, no. 507 (2005): 324–341; S. Nickell and J. Saleheen, *The impact of immigration on occupational wages: British evidence*, SERC Discussion Paper, No. 34, 2008.

Monopoly and monopsony

The theory discussed so far can be amended to consider the cases where a firm has *monopoly power* in its output market (a downward-sloping demand curve for its product) or **monopsony power** in its input markets (an upward-sloping supply curve for its inputs: the firm must then offer a higher factor price to attract a larger quantity of that input).

A firm with **monopsony power** faces an upward-sloping factor supply curve and must offer a higher factor price to attract more factors of production. In expanding inputs, the firm bids up the price paid on all inputs already employed.

The **marginal revenue product of labour (MRPL)** is the change in total output revenue when a firm sells the extra goods that an extra unit of labour input allows it to produce.

Consider a big factory located close to a small town. It is likely that most of the workers of the small town will work for that big factory. The big factory has monopsony power in the local labour market.

For a perfectly competitive firm, the *MVPL* schedule is its marginal revenue from an extra worker. We use the term *marginal value product of labour (MVPL)* for competitive firms who are price-takers in their output markets. *MVPL* is simply the marginal product of labour in physical goods *MPL* multiplied by the output price. We reserve the term **marginal revenue product of labour (MRPL)** for firms with a downward-sloping demand curve for their output.

To find $MRPL$, we use the marginal physical product of labour MPL to work out the extra quantity of output when an extra worker is hired, then calculate the change in the firm's total revenue when it sells these extra goods.

Figure 10.7 shows the $MRPL$ schedules for a monopsony. The $MRPL$ schedule slopes down more steeply than the $MVPL$ schedule in the competitive case because the firm faces a downward-sloping demand curve for its output and recognizes that additional output reduces the price and hence the revenue earned on previous units of output.

The monopsony, in contrast to a competitive firm in the labour market, does not take the wage as given. The wage is now upward-sloping since the monopsonist faces an upward-sloping labour supply. The higher the wage, the higher is the number of workers willing to work. The average cost of the monopsonist is just the wage and it is increasing with the number of workers.

The marginal cost of a monopsony is upward-sloping. A monopsonist recognizes that expanding employment bids up the wage. If all workers are paid the same wage, the marginal cost of an extra worker is not just the wage paid to that worker but also the rise in the wage bill for previously employed workers. The monopsonist's marginal cost of labour exceeds the wage, and rises with the level of employment. This implies that the marginal cost is always above the average cost (ACL). This is shown in Figure 10.7.

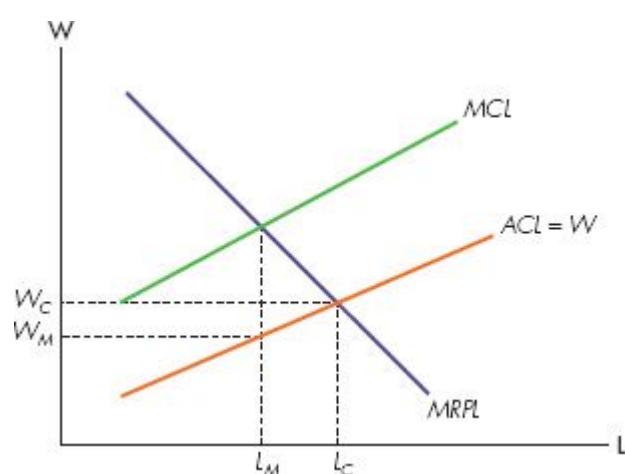


Figure 10.7 Monopsony power

The monopsonist chooses the level of employment that maximizes profits. This happens where $MCL = MRPL$. So the monopsonist chooses a level of employment L_M . The wage paid by the monopsonist is found by looking at the wage curve. The monopsonist pays a wage given by W_M . If, instead of a monopsonist, we have a perfectly competitive labour market, the equilibrium wage will be given by the intersection of the labour supply (the ACL curve) and labour demand (the $MRPL$ curve). In that case, the level of employment will be L_C and the equilibrium wage will be W_C .

Compared to perfect competition, in a monopsony fewer workers are employed. Moreover, the wage paid by the monopsonist is lower than the wage that would be paid in a perfectly competitive labour market.

So far we have assumed that the monopsonist pays the same wage to all its workers. This is the case of a *nondiscriminating* monopsonist. If the monopsonist can *perfectly discriminate*, it pays all its workers different wages. In particular, it pays all its workers their reservation wages. In this case, the monopsonist would choose the same level of employment as in a competitive market.⁴ A perfectly discriminating monopsonist will hire up to the point where the last worker's contribution to its revenue (given by the $MRPL$ curve) equals the marginal cost of labour (W_C). In this case, W_C is not the competitive equilibrium but just the wage paid to attract the last worker.

CONCEPT 10.1

HOW COMMON IS MONOPSONY?

Economists have often assumed that small firms probably face a pretty horizontal labour supply curve – they can attract extra workers without bidding the wage up much. If so, monopsony is more of a special case for textbook writers than something to worry about much in the real world.

However, in the past decade this view has been increasingly challenged. Even small firms not requiring very many extra workers may have difficulty in attracting the workers they need without bidding up the wage they have to offer.

For example, Professors Alan Manning and Steve Machin (2002) studied residential care homes in southern England. Towns like Bournemouth and Eastbourne are famous as places in which the elderly cluster in their retirement. Manning and Machin collected data on the wages paid to individual care workers in individual retirement homes and discovered a surprising fact. There is a very large wage dispersion across care homes, even after controlling for identifiable differences in their workers. This is difficult to reconcile with a labour market in which each firm is a price-taker for labour. Monopsony may be more relevant than you first thought.

Source: adapted from S. Machin and A. Manning, *The structure of wages in what should be a competitive labour market*, Centre for Economic Performance, London School of Economics, 2002.

10.6 Transfer earnings and economic rents

In some sectors workers are paid much more, on average, than in other sectors. For example, talented pianists or footballers generally earn high wages.⁵ Why do they get paid so much? To answer this question, we need to introduce the idea of **economic rent**. The economic rent is the amount paid to a factor of production in excess of what is required to keep that factor in the current occupation. For a worker, the lowest payment needed to keep her in a given occupation is her **reservation wage** for that occupation.⁶ Any payment in excess of her reservation wage represents an economic rent for the worker.

Economic rent (not to be confused with income from renting out property) is the payment a factor receives in excess of what is needed to keep it in its present occupation.

The **reservation wage** is the lowest wage a worker is willing to accept to work in a given occupation.

For example, suppose a worker gets £10 an hour to work in occupation *A*. Her best alternative is to work in occupation *B* at £8.⁷ For this worker, £2 an hour is the part of her wage that represents her economic rent for working in occupation *A*. To stay in occupation *A*, she has to receive at

least £8 an hour. This is her reservation wage for working in occupation A . This represents the opportunity cost of working in occupation A .

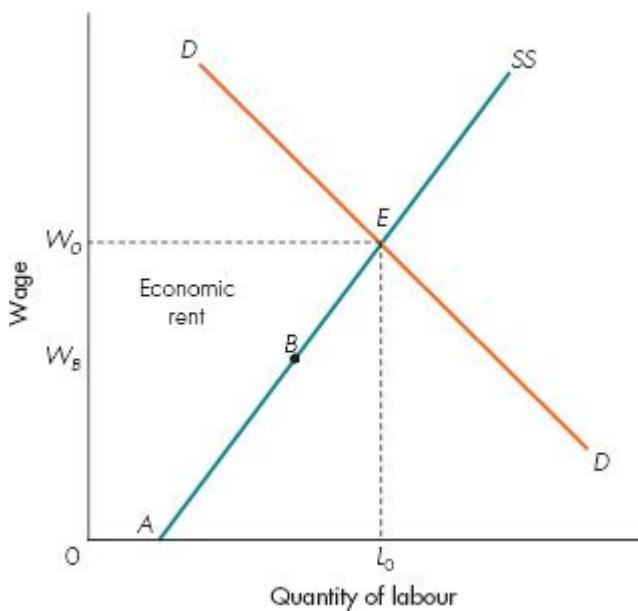
To see how economic rent is determined, let's consider the market for concert pianists. We assume that all workers hired by a firm must be paid the same wage. This is not always true. Nevertheless, our assumption will help us in understanding the basic difference between transfer earnings and economic rents.

In Figure 10.8 DD is the labour demand for concert pianists and SS the supply of pianists to the music industry. Even at a zero wage some dedicated musicians would be concert pianists. The first A pianists would work for nothing. Higher wages attract into the industry concert pianists who could have done other things. A wage W_1 is needed to expand the supply of pianists to B , and W_0 must be paid to increase supply to L_0 . The supply curve slopes upwards.

Because all workers are paid the same wage, equilibrium is at E , with a wage W_0 and a number of pianists L_0 . W_0 may be a large wage. Each firm in the music industry pays W_0 because their workers are very talented, with a high marginal product. In the output market (concerts), firms earn large revenues. The derived demand curve DD for concert pianists is very high.

At the equilibrium E , the last pianist entering the industry has a reservation wage W_0 . This last pianist's marginal value product is also W_0 , since E is on the demand curve DD . Since the industry has to pay all workers the same wage, all previous workers are paid W_0 , even though the labour supply curve SS implies they would have worked for less than W_0 . These workers, with reservation wages below W_0 , earn *economic rent*, a pure surplus arising because W_0 is needed to attract the last pianist. Rent reflects differences in pianists' *reservation wages*, not their *productivity* as musicians.

In Figure 10.8 the industry makes total wage payments equal to the rectangle OW_0EL_0 . It pays L_0 workers W_0 each. The economic rent that workers as a whole receive is $OAEW_0$.



DD is the industry labour demand. A quantity A of labour would work in the industry even at a zero wage. Higher wages attract additional workers to the industry. SS is the industry labour supply. If each worker was paid only his reservation wage, the industry needs only to pay AL_0E in wages and no workers receive economic rents. If all workers must be paid the highest wage rate necessary to attract the last worker to the industry, equilibrium at E implies that workers as a whole derive economic rent $0AEW_0$.

Figure 10.8 Economic rent in the labour market

Economic rent depends on the elasticity of labour supply. If labour supply is totally inelastic (that is, vertical), all payment received by the workers will be economic rent. If labour supply is completely elastic (that is, horizontal), the workers are paid just their reservation wage and there is no economic rent.

Note the distinction between the firm and the industry. Economic rent is an unnecessary payment as far as the industry is concerned. By colluding in order to wage-discriminate, paying each worker his reservation wage alone, the industry could retain all its workers without paying them economic rent. But for a single competitive firm, the wage W_0 has to be paid to keep its workers. If it fails to pay the going rate, its workers will go to another firm.

In the UK football industry and the US baseball industry, it is often said that high player salaries are bankrupting the industry. But wages are high because the derived demand is high – crowds at the ground and

television rights make it profitable to supply this output – and because the supply of talented players is scarce. The supply curve of good players is inelastic: even very high wages cannot increase the number of good players by much. A large proportion of the wages received by professional footballers is economic rent.

CASE 10.3

HIGHER EDUCATION PAYS OFF

Nowadays most students have to contribute to the cost of their higher education. Since October 2012 students in the UK have faced higher university tuition fees. What can we tell the students about the financial benefits they are likely to accrue in the future? The table below shows the results of a major empirical study on determinants of people's wages by the time they are 33 years old.

The research suggests that degrees add a lot to future earning power. This is consistent with the idea of human capital. Human capital is the result of past investment in order to raise future income. The cost of investing in another year of school education or a further qualification is the direct cost, such as tuition fees, plus the opportunity cost of the time involved, namely delaying paid employment. The benefit of the investment is a higher future monetary income or a future job yielding more job satisfaction.

The results in the table also outline another important fact. Investing in higher education by students seems to pay off in terms of future earnings, but the subject studied also matters. Economics students can expect to earn much more than history or language students.

	% extra wage in Britain at age 33 for	
	Men	Women
First degree	+15	+32
Postgraduate degree	+15	+35
Extra effect by subject		
Arts	-10	+5
Economics	+10	+24
Chemistry/biology	-17	-11
Maths/physics	+9	+16



© ericsphotography | istockphoto.com

Source: R. Blundell et al., 'Returns to higher education in Britain', *Economic Journal* 110, no. 461 (2000): 82–99.

10.7 Do labour markets clear?

So far we have assumed that wages are flexible. The equilibrium wage equates labour supply and labour demand. In Part Four you will see that many questions in macroeconomics turn on whether wage flexibility is sufficient to keep labour markets near their equilibrium positions. It may not be possible to take labour market equilibrium for granted.

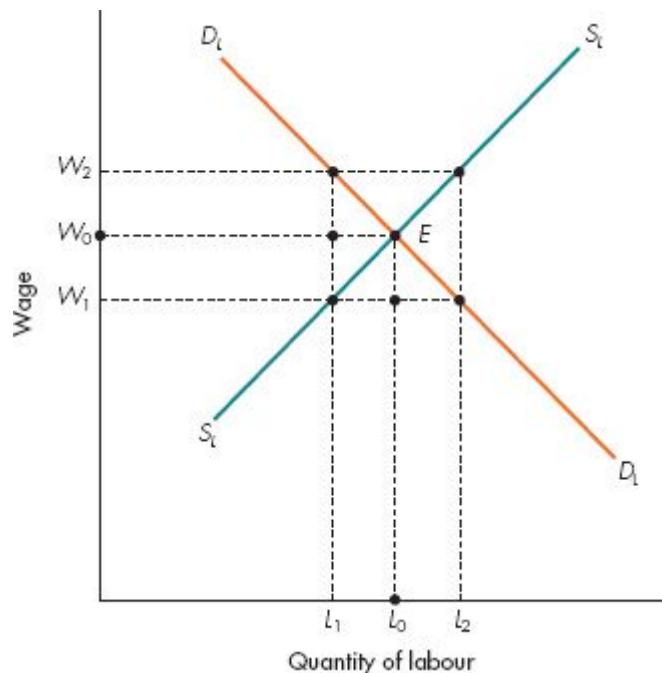
Minimum wage agreements

In 2012 the UK minimum wage was £6.19 an hour. Figure 10.9 shows the demand curve $D_L D_L$ and the supply curve $S_L S_L$ for a particular skill in a particular industry. Free market equilibrium is at E . For skilled workers, the equilibrium wage W_0 exceeds a minimum wage at W_1 , which is thus irrelevant.

Workers are **involuntarily unemployed** if they would work at the going wage but cannot find jobs.

Suppose the minimum wage is W_2 , above the free market equilibrium wage W_0 . At W_2 there is excess labour supply $L_2 - L_1$. Since firms cannot be forced to hire workers they do not want, employment is L_1 and the quantity of workers $L_2 - L_1$ is **involuntarily unemployed**.⁸

A national minimum wage may exceed the free market equilibrium wage for lower-skilled occupations. If so, those workers lucky enough to find jobs get higher wages than before but the total amount of employment is lower than in free market equilibrium. Minimum wages may explain involuntary unemployment among lower-skilled workers.



Free market equilibrium occurs at the wage W_0 and a quantity of employment L_0 . A minimum wage W_1 below W_0 is irrelevant. However, a minimum wage W_2 above W_0 will restrict the actual quantity of employment to L_1 , leaving a quantity $L_2 - L_1$ of workers involuntarily unemployed. They would like to work at this wage rate but cannot find jobs.

Figure 10.9 A minimum wage

CONCEPT 10.2

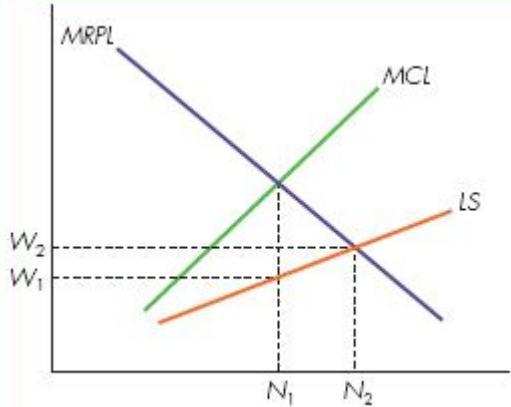
MINIMUM WAGES HURT JOBS, DON'T THEY?

A minimum wage prices some workers out of a job: by raising wages, it slides firms up their demand curves, cutting jobs. Even politicians understand. Right?

The ‘proof’ relies on a competitive labour market. People’s intuition is often based on perfect competition. What happens if there is a sole employer? A monopsonist’s new hiring bids up the price of existing workers: the marginal cost of labour exceeds the wage. The figure below shows the marginal revenue product of labour, the labour supply curve facing the firm, and the marginal cost of labour to the monopsonist. In equilibrium, $MRPL=MCL$. Employment is N_1 and a wage W_1 is needed to attract this labour. The vertical gap between LS and $MRPL$ shows workers are paid less than their marginal product. This is called exploitation.

At a minimum wage W_2 , the monopsonist faces a horizontal labour supply at W_2 , at least until N_2 people are hired. W_2 is now the marginal cost of labour. The firm hires N_2 workers to equate the marginal cost and marginal benefit of hiring. By offsetting exploitation, the minimum wage boosts jobs from N_1 to N_2 .

Beginning at free market equilibrium at a wage W_1 , successive rises in the minimum wage boost jobs (sliding the firm along the labour supply curve LS) until the minimum wage reaches W_2 , at which employment is maximized. Still higher minimum wages now move the firm up its demand curve, reducing jobs thereafter. When firms have some monopsony power, a minimum wage slightly above the free market equilibrium is good for jobs – it offsets the distortion caused by the market power of employers – but a minimum wage substantially above the free market equilibrium is bad for jobs.



Source: J. Dolado et al., 'The economic impact of minimum wages in Europe', *Economic Policy* 11, no. 23 (1996): 319–372.

Trade unions

Trade unions are worker organizations set up by workers to affect pay and working conditions. Do unions protect workers from exploitation by powerful employers or do they use their power to secure unjustified pay increases and oppose technical change and productivity improvements that might threaten the jobs of their members?

In 1980 half the civilian labour force in the UK belonged to a trade union. Figure 10.10 shows changes since 1910. After a steady increase in union membership until the late 1920s there was a massive decline during the Depression of the 1930s. After a sharp recovery until 1950, the degree of unionization of the labour force remained fairly constant until the late 1960s. The 1970s saw a sharp rise in union membership, which peaked in 1979, since when it has been falling sharply.

The trend outlined in Figure 10.10, to different degrees, is common to many other European countries, as we can see in Table 10.4.

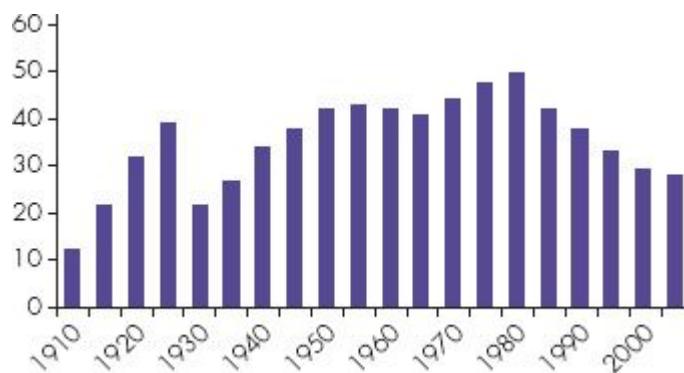


Figure 10.10 Union membership (% of civilian labour force)

Source: Bain, G. S. and Elsheik, F. (1976) *Union Growth and the Business Cycle*, Basil Blackwell; ONS, *Labour Market Trends*.

Table 10.4 Trade union density rates in selected countries (% of civilian labour force)

Country	1980	1990	2000	2011
Sweden	78.2	80.0	79.1	68.0
Denmark	78.6	75.3	74.2	68.8
Italy	54.5	38.8	34.8	35.1
Ireland	57.4	51.0	37.1	35.5
France	17.1	10.1	8	7.5
Germany	34.9	31.2	24.5	18.4

Sources: J. Waddington, *Trade Union Membership in Europe*, 2005 (<http://library.fes.de/pdf-files/gurn/00287.pdf>); OECD.

Declining unionization reflects several trends. First, as the share of the service sector in national output rises, the share in traditional industry, manual and male, has shrunk. Second, the public sector, in which unions were traditionally well organized, has shrunk as a result of privatization and cuts in its size. Third, technical advances have made production much more flexible and small scale, circumstances in which it is harder to organize a trade union. Fourth, increased female participation in the labour force has often been in part-time jobs in which union organization is harder. Finally, globalization has had an impact. Greater international competition in more and more industries is eliminating pockets of domestic monopoly whose profits were tempting targets for unions. As

greater competition makes the derived demand for labour more elastic, unions face an ever-worsening trade-off between wages and employment. Restricting labour supply only raises wages by sacrificing many jobs. As the trade-off gets less attractive, belonging to a union becomes less worthwhile.

The traditional view of unions is that they offset the power that a firm enjoys in negotiating wages and working conditions. Consider a single firm with many workers. If each worker must make a separate deal with the firm, the firm can make a take-it-or-leave-it offer. A worker with firm-specific human capital, which will be almost useless in any other firm, may face a large drop in productivity and wages if he rejects the firm's offer. The firm is in a strong bargaining position if it can make separate agreements with each worker. In contrast, by presenting a united front, the workers may be able to impose large costs on the firm if they *all* quit. The firm can replace one worker but not its whole labour force. The existence of unions evens up the bargaining process.

A **closed shop** is an agreement that all a firm's workers will be members of a trade union.

Once a union is established, it aims not merely to protect its members but also to improve their pay and conditions. To be successful, the union must be able to restrict the firm's labour supply. If the firm can hire non-union labour, unions will find it hard to maintain the wage above the level at which the firm can hire non-union workers. This is one reason why unions are keen on **closed-shop** agreements with individual firms.⁹

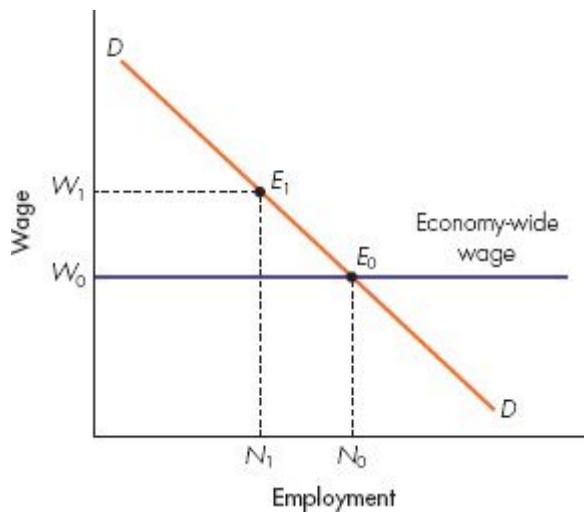
How do unions raise wages by restricting supply? Figure 10.11 shows an industry's downward-sloping labour demand curve DD . The wage in the rest of the economy is W_0 , and we assume the industry faces a perfectly elastic labour supply curve at the wage rate W_0 . In the absence of unions, equilibrium is at E_0 with employment N_0 .

Now suppose everyone in the industry must belong to a trade union and the union restricts labour in this industry to N_1 . The industry faces a vertical labour supply curve at N_1 . Equilibrium is at E_1 . By sacrificing employment in the industry, the union raises the wage for each employed member from W_0 to W_1 . At a higher wage and marginal cost of production, each firm is forced to raise its price. The full effect of the

trade union is not merely to raise wages and lower employment in the industry but also to raise the output price and lower equilibrium output of the industry.

This analysis raises two questions. What determines how far the union will trade off lower employment for higher wages in the industry? And what determines how much power unions have to control the supply of labour to particular industries?

Assume that the union has full control over the supply of labour to a firm or an industry. It can trade off employment for wage rises. How far it will go depends on the preferences or tastes of the union and its members. It might try to maximize total income (wage multiplied by employment) of its members, or it might try to maximize per capita income (wages) of those in employment. A lot depends on the power and decision structure within the union.



In the absence of a union, the industry would face a horizontal labour supply curve at the wage W_0 . Given the industry demand curve DD for labour, equilibrium would occur at E_0 . By restricting the industry labour supply to N_1 , the union can increase the wage to W_1 . It can trade off lower employment in the industry for higher wages.

Figure 10.11 Unions in the labour market

The more the union cares about its senior members, the more it is likely to maximize the wage independently of what happens to employment. Senior workers have the most firm-specific human capital and are the least likely to be sacked if total employment in the industry must fall.

Conversely, the more the union is democratic, and the more it cares about its potential members as well as those actually in employment, the less likely it is to restrict employment to ensure higher wages for those who remain employed in the industry.

Scale economies

Involuntary unemployment may reflect scale economies and imperfect competition. These create entry barriers and prevent new firms from joining an industry. Entry barriers prevent the unemployed from starting new firms even if unemployed workers would work for a lower wage than that paid in existing firms.

Insiders and outsiders

The previous explanation emphasizes entry barriers in forming new firms. Insider–outsider theories emphasize barriers to entering employment in existing firms.

Entry barriers take many forms. It is costly to advertise for workers, interview them, evaluate what sort of job they should be offered and train them in activities specific to the firm. In the terminology of Chapter 9, these are innocent entry barriers.

Insiders have jobs and are represented in wage bargaining.

Outsiders do not have jobs and are unrepresented in wage bargaining.

But existing workers (insiders) may also erect strategic barriers to entry by outsiders, even without the presence of formal trade unions. For example, **insiders** may threaten industrial disruption if too many outsiders are admitted too quickly or if **outsiders** offer to work at a lower wage than that being paid to insiders.

When such entry barriers confront outsiders, the insiders can raise their own wage above that for which outsiders would be prepared to work *without* inducing a spate of hiring of outsiders.

Efficiency wages

Thus far, we have assumed that information is cheap to come by. In practice, firms face two problems: it is hard to tell whether a job applicant will be a productive worker (a matter of innate ability) and hard to monitor whether workers shirk after being employed.¹⁰

Efficiency wages are high wages that raise productivity through their incentive effect.

Given the cost of evaluating new workers, and the subsequent cost of monitoring their performance on the job, what is the best policy for a firm? The **efficiency wage** theory argues that it is profitable for firms to pay existing workers a wage above their reservation wages.

First, suppose workers quit their job if they get a better offer elsewhere. If firms pay a wage that is the average of that faced by productive and unproductive workers, it is the productive workers who are more likely to find better offers elsewhere and quit. Eventually, the firm will be left only with the low-quality workers. Paying a wage premium helps retain high-quality workers, even if the firm has some trouble telling which these are.

Second, when workers shirk on the job they may get caught. If caught, they get sacked. How big is the penalty for being caught? It is the difference between the current wage and what the worker gets in unemployment benefit or in a subsequent job. The higher the wage paid by the existing employer, the larger the penalty of being caught shirking. To increase the penalty and reduce the incentive to shirk, firms pay existing workers a higher wage than on average is necessary to get them to supply their labour.

The implication is that some workers may be involuntarily unemployed. They may be happy to work for wages at or below those paid to existing workers but have little practical chance of securing a job at such wages.

Minimum wage agreements, trade union power, scale economies, insider–outsider distinctions and efficiency wages are *possible* explanations for insufficient wage flexibility in the short run to maintain the labour market in continuous equilibrium. Whether the labour market is always in equilibrium, and the length of time for which disequilibrium persists, are questions to which we return repeatedly in Part Four.

10.8 Wage discrimination

Wage discrimination refers to a situation whereby equally productive workers are paid differently simply because they differ in some personal characteristic, such as gender, race or age. There are two main explanations for wage discrimination. The first explanation is known as *taste-based discrimination*.¹¹ According to this approach, employers may have a preference for discrimination, meaning they experience a disutility as a result of hiring a worker from a minority group.

Suppose there are two groups of workers in the economy, majority and minority workers. A is the group of majority workers and B is the group of minority workers. Workers in both groups are equally productive and are in equal number. Suppose a monopsonist has a prejudice towards workers coming from a minority group.

$MRPL_A$ is the marginal revenue product of majority workers and $MRPL_B$ is that of minority workers. Since all workers have the same productivity, the marginal revenue product of the two groups is the same. This situation is shown in Figure 10.12.

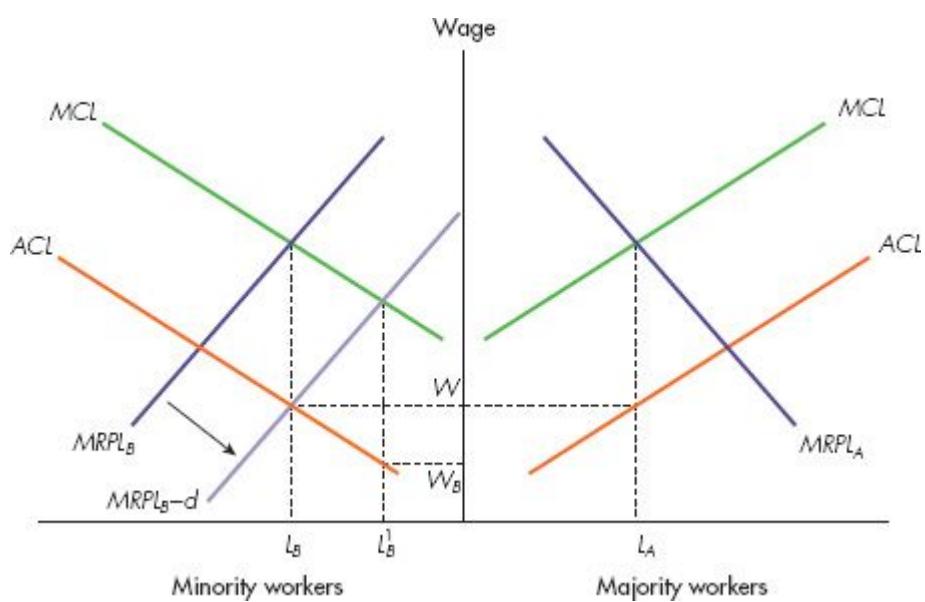


Figure 10.12 A discriminating monopsonist

The monopsonist maximizes profits and hires at the point where the $MRPL$ of workers is equal to the marginal cost. If the monopsonist has no taste for discrimination, it will hire both types of workers, offering them the same wage W . It will hire L_B minority workers and L_A majority workers.

If the monopsonist has a taste for discrimination, it views the $MRPL$ of minority workers as being of less value than that of majority workers by a factor d , which we call the discrimination factor. Therefore the relevant marginal revenue product of minority workers becomes $MRPL_B / d$.

Now it will hire only L_B^1 minority workers. There will be a wage gap between the two groups. Minority workers receive a wage W_B , which is lower than the wage received by majority workers (W). Minority workers have to compensate the employer by accepting a lower wage or being more productive than majority workers at a given wage.

The second explanation for wage discrimination is known as *statistical discrimination*. The idea behind this approach is that employers may have limited information about the skills of job applicants. In this case, employers may observe personal characteristics of job applicants, like race or gender, to infer the productivity of those applicants. For example, suppose that an employer believes that male workers are on average more productive than female workers. Then the employer is more likely to hire male rather than female applicants.

Employment discrimination on the basis of personal characteristics of workers is normally illegal. Nevertheless some cases of discrimination may be difficult to detect so it is plausible that, despite being unlawful, they occur more frequently than we think.

The wage gap between men and women

Few women do manufacturing jobs; most have jobs in services. Yet the pattern of employment is not the major reason why women on average earn 80 per cent as much as men. Sector by sector, women systematically get paid substantially less than men.

The percentage of women in professional or managerial occupations is comparable with that for men but few women are on the boards of major companies. Why do firms promote or train women more slowly? Suppose firms bear some of the cost of training. Assuming men and

women are of inherently equal ability and educational attainment, it costs the firm the same to train either sex.

Suppose firms believe women are more likely than men to interrupt, or even end, their careers at a young age. As a matter of biology, women have babies. Firms may conclude that the extra productivity benefits in the future are lower for women than men simply because many women work fewer years in the future. It is more profitable to train and promote men.

Some women plan to have a full-time career, either remaining childless or returning to work almost immediately after any children are born. It would make sense for firms to invest in such people. How is a firm to tell which young women are planning to stay? Asking is pointless. There is no incentive for young women to tell the truth.

Suppose firms offer young workers the choice between a relatively flat age–earnings profile and a steep profile that begins at a lower wage but pays a much higher wage later in a worker’s career. In this way, firms can make the two profiles of equal value to someone planning a lifetime career. The early sacrifice (low wages) is recouped with interest later (high wages). Someone planning to quit the labour force, say at the age of 30, will never opt for the steeper profile.

Age–earnings profiles may induce recruits to reveal their true career plans. If women, or any other group with a high risk of quitting at a young age, accept the steeper profile, the firm can embark on training with some confidence that its investment will not be wasted.

Some firms may still try to pay female workers less than male workers who are identical in every respect. This is wage discrimination, as previously discussed.

Our analysis suggests that paternity leave for fathers, the provision of crèches for working parents or a greater acceptance of part-time working by both sexes would increase the incentive for firms to decide to favour the training of women.

In Table 10.5 some data about the wage differential between men and women for some European countries are reported. The wage differential between men and women is measured as a percentage of the difference between the median wages of men and women relative to the median wage of men. In the UK, in 2007, women tended to earn a wage 21 per cent lower than men.

Table 10.5 Wage gap between men and women

Country	Gender wage gap (%)	
	1997	2007
Denmark	13	9
France	10	12
Germany	24	23
Spain	29	17
Sweden	17	15
UK	26	21

Source: OECD, *Employment Outlook*, 2009.

In most of the countries the wage gap between men and women has decreased over time; nevertheless, it is still substantial in many cases.

Summary

- In the long run, a firm chooses a **production technique** to minimize the cost of a particular output. By considering each output, it constructs a **total cost curve**.
- In the long run, a **rise in the price of labour** has a **substitution effect** and an **output effect**. The substitution effect reduces the quantity of labour demanded as the capital–labour ratio rises at each output. But total costs and marginal costs of output increase. The more elastic the firm's demand curve and marginal revenue curve, the more the higher marginal cost curve reduces output, reducing demand for factors. For a higher price of a factor, the substitution and output effects both reduce the quantity demanded.
- In the short run, the firm has **fixed factors**, and probably a **fixed production technique**. The firm can vary short-run output by varying its variable input, labour, which is subject to diminishing returns when other factors are fixed. The **marginal physical product of labour** falls as more labour is hired.

- A profit-maximizing firm produces the output at which marginal cost equals marginal revenue. Equivalently, it hires labour until the **marginal cost of labour** equals its **marginal revenue product**. One implies the other. If the firm is a price-taker in its output market, the *MVPL* is its **marginal value product**, the output price multiplied by its marginal physical product. If the firm is a price-taker in the labour market, the marginal cost of labour is the wage rate. A perfectly competitive firm equates the real wage to the marginal physical product of labour.
- The downward-sloping marginal physical product of labour schedule is the **short-run demand curve for labour** (in terms of the real wage) for a competitive firm. Equivalently, the marginal value product of labour schedule is the demand curve in terms of the nominal wage. The *MVPL* schedule for a firm shifts up if the output price increases, the capital stock increases or if technical progress makes labour more productive.
- The **industry's labour demand curve** is not merely the horizontal sum of firms' *MVPL* curves. Higher industry output in response to a wage reduction also reduces the output price. The industry labour demand curve is steeper (less elastic) than that of each firm, and more inelastic the more inelastic is the demand curve for the industry's output.
- Labour demand curves are **derived demands**. A shift in the output demand curve for the industry will shift the derived factor demand curve in the same direction.
- For someone already in the labour force, a **rise in the hourly real wage** has both a **substitution effect** tending to increase the supply of hours worked, and an **income effect** tending to reduce the supply of hours worked. For men, the two effects cancel out almost exactly in practice but the empirical evidence suggests that the substitution effect dominates for women. Thus women have a rising labour supply curve; for men, it is almost vertical.

- Individuals with non-labour income may prefer not to work. Four things raise the **participation rate in the labour force**: higher real wage rates, lower fixed costs of working, lower non-labour income and changes in tastes in favour of working. These explain the trend for increasing labour force participation by married women over the past few decades.
- The **industry supply curve of labour** depends on the wage paid relative to wages in other industries using similar skills. **Equilibrium wage differentials** are the monetary compensation for differences in non-monetary characteristics of jobs in different industries undertaken by workers with the same skill.
- When the labour supply curve to an industry is less than perfectly elastic, the industry pays higher wages to expand employment. For workers prepared to work in the industry at a lower wage, there is an element of **economic rent** (the difference between income received and the reservation wage for that individual).
- In free market equilibrium, some workers choose not to work at the equilibrium wage rate. They are **voluntarily unemployed**. **Involuntary unemployment** is the difference between desired supply and desired demand at a disequilibrium wage rate. Workers would like to work but cannot find a job.
- There is considerable disagreement about how quickly labour markets can get back to equilibrium if initially in disequilibrium. Possible causes of involuntary unemployment are **minimum wage agreements**, **trade unions**, **scale economies**, **insider–outsider** distinctions and **efficiency wages**.

Review questions

 connect

EASY

- 1 (a) Explain why the marginal product of labour eventually declines. (b) Show in a diagram the effect of an increase in the firm's capital stock on its demand curve for labour.
- 2 (a) Over the past 100 years the real wage has risen but the length of the working week has fallen. Explain this result using income and substitution effects. (b) Explain how an increase in the real wage could cause everyone in employment to work fewer hours but still increase the total amount of work done in the economy.
- 3 Why should the labour supply curve to an industry slope upwards even if the aggregate labour supply to the economy is fixed?
- 4 **Common fallacies** Why are the following statements wrong? (a) There is no economic reason why a sketch that took Picasso one minute to draw should fetch £100 000. (b) Higher wages must raise the incentive to work.
- 5 (a) Why can a top golfer earn more in a weekend than a university professor earns in a year? (b) Why can students studying economics expect to earn more than equally smart students studying philosophy?

MEDIUM

- 6 The labour supply and the labour demand in a competitive labour market are reported in the following table:

L	W^D	W^S
1	15.5	5
2	15	6
3	14.5	7
4	14	8
5	13.5	9
6	13	10
7	12.5	11
8	12	12
9	11.5	13

W^D is the inverse labour demand and W^S is the inverse labour supply.

In a graph with the wage W on the vertical axis and labour L on the horizontal axis, show the labour market equilibrium. Suppose that

labour demand comes from many identical perfectly competitive firms.

If the price of the output produced by those firms is reduced by half because of a recession, explain what happens to labour demand. How will the equilibrium of the labour market be affected?

- 7 Suppose that the labour supply of workers is $L^S = w - 5$ and the labour demand is $L = 20 - w$. Derive the labour market equilibrium.
- 8 Could a university degree increase your subsequent job prospects, even if the subject that you studied at university had no relevance whatsoever to your subsequent career?
- 9 Consider the market for theatre actors where some actors are paid high wages and some are paid low wages. Some actors are ready to work even at zero wages. Show on a graph what happens if all the actors must be paid the highest wage rate necessary to attract more actors into the industry. What is the economic rent earned by the actors?
- 10 ‘A minimum wage set sufficiently high will always reduce jobs, but whether a modest level of minimum wage reduces or increases employment depends entirely on the degree of competition in the labour market.’ Explain.
- 11 Suppose there are two groups of workers in the economy, male workers and female workers. Workers in both groups are equally productive and are in equal number. Suppose a monopsony has a prejudice towards female workers. Show on a graph the effect of this discrimination on the wages the two groups receive in the monopsony.

HARD

- 12 A firm is producing chocolate bars using only labour. The production function is $Q = 20L - 0.5L^2$, where L denotes labour. The firm is selling its chocolate bars in a competitive market and the price of a chocolate bar is £1. The firm hires workers in a competitive market. The wage paid to a worker is w . Write down the profit function of the firm and find the labour demand function (L as a function of w) of the firm.
- 13 **Essay question** In the past 50 years, there has been a dramatic increase in female participation in the labour force. Three possible explanations are: (a) a change in social attitudes toward women working, (b) technological advances that make it easier to accomplish household chores (shopping, cleaning, and so on) without women themselves having to remain at home full time, and (c) the possibility that material

goods are a luxury and that people wish to buy disproportionately more of them as living standards rise. What evidence would you gather in order to test these different hypotheses?

- | 4 Suppose a monopsonist faces the production function $Q = 20L - 0.5L^2$ and a labour supply $L^S = w - 5$. This means that the wage that the monopsony must pay is $w = L + 5$. Assume that the price of the product is £1. Find the labour demand that maximizes the profits of the monopsony. What about the wage?
-

- 1 In 2011 the average hourly real wage in the UK was £12.62. The same wage in 1986 (inflation adjusted) was £7.78. In 25 years hourly real earnings increased by 62 per cent ('Real wages up 62% on average over the past 25 years', ONS, 2012).
- 2 This conclusion applies to small changes in real wage rates. In most Western countries, the large rise in real wages over the past 100 years has been matched by reductions of ten hours or more in the working week.
- 3 Even if some workers in the industry have a backward-bending labour supply curve, the industry labour supply curve is upward sloping. A wage increase in the industry not only increases the supply of labour of some of the existing workers, it also attracts workers from other industries.
- 4 Perfect discrimination is a different concept to wage discrimination, which is analysed in Section 10.8.
- 5 According to the *Daily Mail*, in 2011 the average Premiership footballer earned £22 353 a week. A staggering £1 160 000 a year!
- 6 The reservation wage as defined above is sometimes called *transfer earning*.
- 7 The best alternative may also be not working at all.
- 8 Involuntary unemployment is distinguished from voluntary unemployment, which happens when workers prefer not to work at the existing wage and to remain unemployed.
- 9 Unions frequently argue that, in the absence of a closed shop, non-union workers will benefit from improvements in pay and conditions achieved through the efforts of the union. Non-union members are getting a 'free ride' without paying their union subscriptions.
- 10 Economists refer to these problems as adverse selection and moral hazard. We discuss them in detail in Chapter 12 when we examine the economics of information.
- 11 This approach was introduced by Gary Becker in 1957 in his seminal work, *The Economics of Discrimination*.

CHAPTER 11

Factor markets and income distribution

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 recognize the markets for capital and land
- 2 understand how incomes of factors are determined
- 3 analyse functional and personal distributions of income
- 4 recognize flows over time and stocks at a point in time
- 5 describe the markets for capital services and for new capital assets
- 6 define the concept of present value
- 7 analyse nominal and real interest rates
- 8 understand how saving and investment determine the real interest rate
- 9 describe how land is allocated between competing uses

The previous chapter focused on labour. We now examine the other inputs with which labour co-operates in production. In particular, we focus on physical capital and land. Some issues can be dealt with briefly. You already know how a firm chooses a production technique in the long run, when all factors can be freely varied, and you are familiar with the concept of a factor's marginal product.

Apart from investment in human capital, many aspects of labour market behaviour are easily analysed within a short-run time horizon. Labour is a variable input even in the short run. Since it takes much longer to adjust other factor inputs, decisions about their use must take a longer view.

One theme of this chapter is how the future affects the present. We examine how decisions today should value future benefits and costs, and show how to discount future payments or receipts to calculate their *present value*.

Our interest in the markets for capital and land goes beyond the equilibrium quantity of capital or the equilibrium price of land. There are two reasons to study factor markets as a whole. First, firms rarely use a single input. Decisions about inputs of capital and land affect the demand curve for labour and the equilibrium wage, just as decisions about labour inputs affect the demand for other factors of production.

Second, having completed our analysis of factor markets, we can discuss what determines the *income distribution*. The price of a factor, multiplied by the quantity used, tells us its income. We need to know the prices and quantities of all productive factors to understand how the economy's total income is distributed. We end this chapter by examining income distribution in the UK.

Physical capital is the stock of produced goods that are inputs to production of other goods and services.

The stock of **physical capital** includes assembly-line machinery used to make cars, railway lines making transport services, school buildings producing educational services, dwellings that produce housing services, and consumer durables, such as televisions, that produce entertainment services.

Physical capital is distinguished from **land** by the fact that the former is produced.

Land is the factor of production that nature supplies.

Although nature can change the quantity of land – for example, through earthquakes – production that nature economists treat land as fixed in supply. Its quantity is largely unaffected by economic sueltes, decisions. The distinction between land and capital can become blurred. Fertilizer and irrigation can ‘produce’ better land. Because land and capital may be hard to disentangle, chaptersame chapter. However, the distinction is often useful.

chapter 6 introduced *depreciation*; that is, the extent to which an asset is used up within the period of analysis. Capital and land are both assets. Capital depreciates a little every year, though new capital can be produced. In treating land as fixed, we assume it does not depreciate.

Capital and land are wealth or assets because they are durable. They are **tangible** because they are physical and we could touch them. Financial wealth is not tangible, and not a physical input to production, though it can hire such inputs. We distinguish between *physical* capital – plant, machinery and buildings, which henceforth we call ‘capital’ – and *financial* capital, or money and paper assets.

Together, capital and land are the **tangible wealth** of the economy.

11.1 Physical capital

Table 11.1 shows the level and composition of net physical capital in the UK in 2009. (Data on capital take ages to collect!) The term ‘net’ means net of depreciation. Dwellings are houses and flats. Productive fixed capital is plant, machinery and buildings. Productive capital that is not fixed is called *working capital*: inventories or stocks of manufactured goods awaiting sale, partially finished goods (work in progress) and raw materials held for future production. Inventories are capital because they are produced goods that contribute to future production.

Table 11.1 UK capital stock, 2009

	£bn	%
Dwellings	1309	41
Productive fixed capital	1873	59
Total capital stock	3182	100

Source: ONS, UK National Accounts.

Table 11.2 shows productive fixed capital (*PKF*) used in production in the UK in 2005 and 2009. The final row shows that, at 2006 prices, the capital per employed worker rose from £54 000 to £62 000 during 2005–09.

Table 11.2 Capital input to UK production

	2005	2009
<i>PKF</i>	1565	1782
PKF/employed worker	54	62

Note: *PKF* in £bn. *PKF per worker* in £000, both at 2006 prices.

Source: ONS, UK National Accounts.

Investment in physical capital increases capital as a production input, not only in absolute terms but also relative to the number of workers employed. Table 11.2 shows that, in less than five years, production techniques have become more *capital intensive*. Each worker has more capital with which to work.

Because capital depreciates over time, it takes some investment in new capital goods merely to keep the level of capital constant. Suppose the hard drive of the computer you use for work breaks down. If you buy a new hard drive with similar features to the old one, you are making an investment that just covers the depreciated capital. If instead you buy a better performing hard drive or even a whole new computer, you make an investment that creates a net addition to the capital stock.

Gross investment is the production of new capital goods and the improvement of existing capital goods.

Net investment is gross investment minus depreciation of the existing capital stock.

Firms undertake investments to make new additions to their capital stock and to replace existing capital that has depreciated. If **net investment** is positive, **gross investment** more than offsets depreciation. The capital stock rises. Conversely, if gross investment fails to offset depreciation, the capital stock falls.

11.2 Rentals, interest rates and asset prices

Table 11.3 distinguishes between *stocks* and *flows* and between *rental payments* and *asset prices*. The price for hiring labour services is the wage. Rather loosely, we call it the ‘price of labour’ but the wage is the *rental payment* to hire labour. There is no asset price for buying the physical asset called a ‘worker’. We no longer have slavery, that is, ownership of workers by firms.

Table 11.3 Stock and flow concepts

	Capital	Labour
Flow input to hourly production	Capital services	Labour services
Payment for flow	Rental rate (£/machine hour)	Wage rate (£/labour hour)
Asset price	£/machine	£/slave, if purchase allowed

The **price of an asset** is the sum for which the asset can be purchased outright. The owner of a capital asset gets the future stream of capital services from this asset.

The cost of using capital services is the **rental rate** for capital.

A **flow** is the stream of services an asset provides in a period of time.

A **stock** is the quantity of an asset at a point in time.

Unlike labour, capital goods can be bought (and sold) and have an **asset price**. A firm’s decision to buy capital goods is an investment decision. The use of capital goods over time provides *capital services* to the owner of this capital. Buying a car for £9000 entitles you to a stream of future transport services. You might even obtain a stream of future rental payments by renting your car to others.

However firms do not have to buy capital goods to get capital services. They can rent capital assets from others and pay them a **rental rate**. The rental rate is the price paid to use the capital asset for a limited period of time. It represents the opportunity cost of using capital. A firm that uses a capital good to produce could have rented it out. By using the capital good instead of renting it, the firm loses

possible rental payments. Renting capital may not always be possible because there may not be a rental market; for example, it is impossible to rent a power station.

Rental payments and asset prices correspond to **flows** and **stocks**. They are different prices but obviously related. When firms make a once-and-for-all purchase of a capital asset they calculate how much it is implicitly costing them to use that capital. A firm is willing to pay more for a capital good if it produces a valuable stream of future rental payments.

In the following we analyse first how firms take investment decisions. Then, starting in Section 11.4, we analyse the decision of firms to supply and demand capital services and show how the market rental rate is determined. In Section 11.7 we discuss how the market for capital goods determines the asset price of capital.

To understand how firms take investment decisions we need to understand what a firm will pay for a capital asset. This reflects the value of the future income from capital services that the capital asset provides. However, we cannot simply add the future rental payments over the life of the capital asset to calculate its current asset price or value. We have to pay attention to the role of *time* and *interest payments*.

Interest and present values

A lender makes a loan to a borrower, who agrees to repay the initial sum (the principal) with *interest* at some future date. A loan of £100 for a year at 10 per cent interest must be repaid at £110 by the end of the year. The extra £10 (10 per cent of £100) is the interest cost of borrowing £100 for a year. *Interest rates* are quoted as a percentage per annum.

Suppose we lend £1 and re-lend the interest as it accrues. The first row of Table 11.4 shows what happens if the annual interest rate is 10 per cent. After a year, we have £1 plus an interest payment of £0.10. Re-lending the whole £1.10, we have £1.21 by the end of the second year. Because of *compound interest* – that is, the process of adding interest rate payments to an original sum of money – the absolute amount by which our money grows increases every year. The first year we increase our money by £0.10, which is 10 per cent of £1. Since we re-lend the interest, our money grows by £0.11 in the next year since we earn 10 per cent on £1.10. If we lend for yet another year, our money will grow by £0.121 to £1.331 at the end of the third year.

Table 11.4 Interest and present value (PV)

	Year		
	0	1	2
<i>At 10% interest rate:</i>			
Value of £1 lent today in:	£1	£1.10	£1.21
PV of £1 earned in:	£1	£0.91	£0.83

	Year		
	0	1	2
<i>At 5% interest rate:</i>			
Value of £1 lent today in:	£1	£1.05	£1.10
PV of £1 earned in:	£1	£0.95	£0.91

At 10 per cent interest per annum, £1 in year 0 is worth £1.10 in year 1 and £1.21 in year 2. Now ask the question the other way round. If we offered you £1.21 in two years' time, what sum today would be just as valuable? The answer is £1. If you had £1 today, you could always lend it out to get exactly £1.21 in two years' time. The second row of Table 11.4 extends this idea. If £1.21 in year 2 is worth £1 today, then £1 in year 2 must be worth £[1/1.21] = £0.83 today; £0.83 today could be lent out at 10 per cent interest to accumulate to £1 in year 2. Similarly, £1 in year 1 is worth only £[1/1.10] = £0.91 today.

Compound interest implies that lending £1 today cumulates to ever-larger sums the further into the future we keep the loan and re-lend the interest. Conversely, the present value of £1 earned at some future date becomes smaller the further into the future the date at which the £1 is earned.

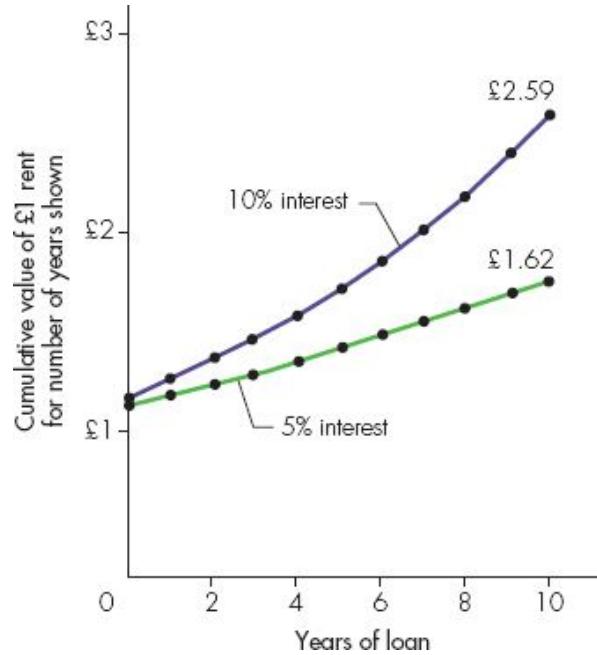
The **present value** of a future £1 is the sum that, if lent today, would cumulate to £1 by that date.

The **present value** of a future payment also depends on the interest rate. Table 11.4 shows that a loan of £1 accumulates less rapidly over time if the interest rate is lower. At 5 per cent interest, a loan of £1 cumulates to only £1.10 after two years, compared with £1.21 after two years when the interest rate was 10 per cent in row 1. Hence the bottom row of Table 11.4 shows that the present value of £1 in year 1 or year 2 is larger when the interest rate is only 5 per cent than in the corresponding entry when the interest rate is 10 per cent.

Figure 11.1 illustrates the same points, showing how lending £1 today cumulates at compound interest rates of 5 and 10 per cent. After 10 years, the loan fund is worth £2.59 at 10 per cent interest but only £1.62 at 5 per cent interest. Higher interest rates imply more rapid accumulation through lending. The same diagram can be used for present values. A payment of £2.59 in 10 years' time has a present value of £1 if the annual interest rate is 10 per cent. The present value of £1 in 10 years' time is thus £1/[2.59] = £0.386. If interest rates are 5 per cent, the value of £1 in 10 years' time is £1/[1.62] = £0.617.

Using interest rates to calculate present values of future payments tells us the right way to add together payments at different points in time. For each payment at each

date, we calculate its present value. Then we add together the present values of the different payments.



At 10 per cent interest per annum, £1 accumulates £2.59 after 10 years. At the lower interest rate of 5 per cent per annum, the accumulated interest value rises much more slowly, reaching only £1.62 after 10 years.

Figure 11.1 Accumulation through interest

To relate the price of a capital asset to the stream of future payments earned from the capital services it provides, we calculate the present value of the rental payment earned by the asset in each year of its working life, and add these present values together. This is what the asset is worth today. In equilibrium, it should be the asset price.

Asset valuation

How much would you bid for a machine that earns £4000 in rental for two years and is then sold for scrap for £10 000? If you bid anything without finding out the interest rate, you misunderstood the previous section! Suppose the annual interest rate is 10 per cent. The first two rows of Table 11.5 show the money received each year. The final column shows the present value of these receipts. From Table 11.4, £1 next year is worth only £0.91 today, and £1 in year 2 only £0.83 today. The present value PV of £4000 in year 1 is £3640 ($\text{£}4000 \times 0.91$), and the PV of the £14 000 received from rental earnings and sale for scrap in year 2 is £11 620 ($\text{£}14\,000 \times 0.83$). Adding these present values for years 1 and 2, the asset price should be £15 260.

Thus, £15 260 is much smaller than the £18 000 actually earned from two years of rental income and the scrap value. Present values *discount* the future.

Table 11.5 Present values and asset prices (at an annual interest rate of 10%)

Year	Rental (£)	Scrap value (£)	Present value (£)
1	4000		3 640
2	4000+	10 000	11 620
Asset price in year 0			15 260

Note: From Table 11.4, the present value of each £1 in year 1 is £0.91 and in year 2 is £0.83 when the interest rate is 10%.

These principles can be used to calculate the present value of any future income stream once the interest rate is known. The calculation is very simple in one special case: when the asset lasts for ever and the income stream per time period is constant. Governments sometimes borrow by selling a *perpetuity*; that is, a bond (simply a piece of paper) promising to pay the owner a constant interest payment (called the ‘coupon’) forever. In the UK, these are called ‘consols’ (after a famous bond issue called Consolidated Stock). The *PV* of a consol – the price the stock market will offer for this piece of paper – obeys the formula

$$PV = \frac{\text{constant annual coupon payment}}{\text{interest rate per annum}} \quad (1)$$

In the financial pages of a newspaper you will find 2.5 per cent consols. This perpetuity promises to pay £2.50 per annum forever. £2.50 was 2.5 per cent of the original sale price of £100. If the current rate of interest is 5 per cent, 2.5 per cent consols should be worth around £50 (the annual coupon £2.5, divided by 0.05, the annual interest rate as a decimal fraction). If interest rates rise to 10 per cent per annum, the consol is then worth only £25 = [(£2.5)/(0.10)].

MATHS 11.1

THE SIMPLE ALGEBRA OF PRESENT VALUES AND DISCOUNTING

Suppose we lend £ K today at an annual interest rate i . After one year, our money has grown to £ $\{K(1 + i)\}$. With $K = 100$ and $i = 0.1$, we get £110 back after a year. If we re-lend the money for another year at the same interest rate, we get back £ $\{K(1 + i)\}(1 + i)$ at the end of the second year. For example, our £100 has grown to £121 after two years. If we lend this sum for yet another year, we get back £ $K(1 + i)^3$ at the end of the third year. Hence, after N years we get back £ $K(1 + i)^N$. This process tells you how to calculate the value after N years of an amount £ K that you have today.

Conversely, the present value of £ X to be received N years later is $\frac{X}{(1+i)^N}$, and we call $\frac{1}{(1+i)^N}$ the *discount factor*. Since the interest rate i is a positive number, the discount factor must be a positive fraction. Higher interest rates imply lower discount factors. The table below shows the present value of £1 N years from now when the interest rate is 10 per cent a year ($i = 0.1$).

Present value (PV) of £1 N years from now, annual interest rate of 10%						
N	1	5	10	20	30	40
PV	£0.91	£0.62	£0.39	£0.15	£0.06	£0.02

To calculate the present value of a whole stream of future payments, we multiply the face value of each payment by the relevant discount factor.

Assuming that the interest rate remains constant, the present value of a future stream of revenues over N years from now is given by:

$$PV = \sum_{t=1}^N \frac{R_t}{(1+i)^t}$$

where R_t is the revenue in year t , i is the interest rate and Σ is a symbol that means the sum of each year's discounted earnings $R_t/(1+i)^t$.

For example, suppose a firm wants to buy today a machine that costs £8000. The machine can give a revenue of £2000 a year for four years. After four years, the machine can be sold as scrap for £3000. Assume that the interest rate is 10 per cent in all four years. The present value of this stream of future revenues is:

$$PV = \frac{2000}{(1+0.1)} + \frac{2000}{(1+0.1)^2} + \frac{2000}{(1+0.1)^3} + \frac{5000}{(1+0.1)^4} = £8388.7$$

In this case, the present value of the future revenues from the machine is higher than the cost of buying the machine. The firm should indeed buy the machine in this case.

The difference between the present value of a stream of revenues from a given investment minus the actual cost of that investment is called the *net present value (NPV)*. In our case, the net present value from buying the machine is $NPV = 8388.7 - 8000 = £388.7$.

This provides a rule for investment decisions: you should invest in a particular project if the net present value of that project is non-negative.

An alternative way to assess whether an investment should be undertaken is given by the calculation of the required real rate of return. This is discussed in Section 11.5.

When an asset is a perpetuity, earning £ K a year for ever, formula (1) implies that the present value of this stream is $£K/i$.

Real and nominal interest rates: inflation and present values

Thus far we have discussed future payments valued in nominal terms. The first column of Table 11.5 shows rental receipts in actual pounds. The interest rate of 10 per cent tells us how many actual pounds we earn by lending £1 for a year.

The **nominal interest rate** tells us how many actual pounds are earned by lending £1 for a year.

The **real interest rate** on a loan is the extra quantity of goods that can be purchased.

At a **nominal interest rate** of 10 per cent, £100 lent today accumulates to £110 by next year. But we want to know how many goods that £110 will then buy. This is what really matters for the lender.

Suppose the nominal interest rate is 10 per cent and inflation is 6 per cent when the lender receives back the money lent. Lending £1 for a year gives £1.10. Since inflation is 6 per cent, it costs £1.06 to buy goods we could have bought for £1 today. With £1.10 to spend next year, our purchasing power rises by only 4 per cent. The **real interest rate** is 4 per cent. Thus

$$\text{Real interest rate} = \text{nominal interest rate} - \text{inflation rate} \quad (2)$$

Consider another example: nominal interest rates are 17 per cent and inflation is 20 per cent. Lending £100 for a year, you get £117. But it will cost you £120 to buy goods you could have bought today for £100. You are worse off by 3 per cent by delaying purchases for a year and lending your money at the apparently high rate of 17 per cent. Real interest rates are *negative*. The real interest rate is —3 per cent. In real terms, it *costs* you to be a lender. The nominal interest rate does not compensate for higher prices of goods you ultimately wish to buy. Notice that the nominal interest rate *cannot* be negative whereas the real interest rate can.

Equation (2) is known as the Fisher equation.¹ It implies that the nominal interest rate can be written as the sum of the real interest rate and the inflation rate. Therefore, for a given real interest rate inflation and nominal interest rate should move together. This is shown in Figure 11.2, where we plot the nominal interest rate (as measured by the monthly average official bank rate) and the monthly inflation rate (measured by the change in the retail price index) in the UK.

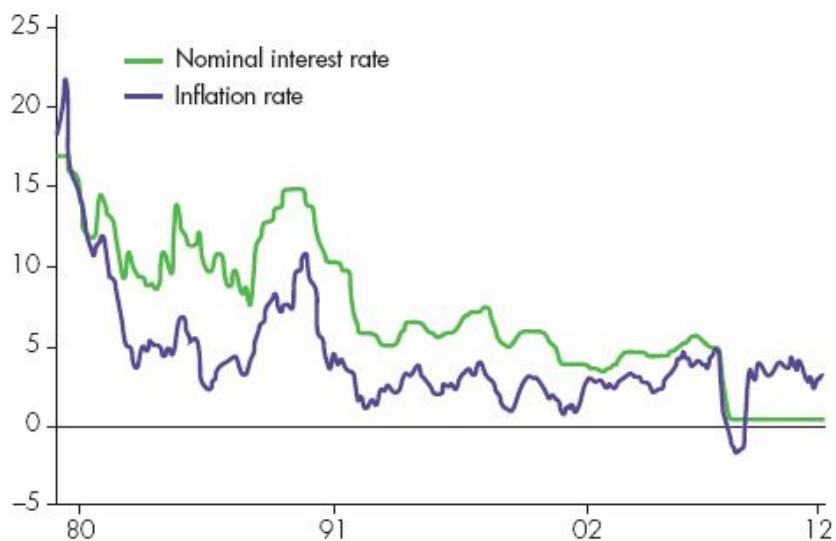


Figure 11.2 Nominal interest rate and inflation rate in UK: monthly averages
 Sources: Bank of England; ONS.

What determines the real interest rate?

Two forces lead to positive real interest rates. First, people are impatient. Given the choice of an equal number of goods tomorrow or today, we'd rather have them today. To delay spending on goods and services, savers have to be compensated with a positive real interest rate that lets them consume *more* goods in the future if they postpone consumption and lend today.

Second, there must be a way of earning positive real returns, or borrowers would never borrow. Borrowers pay positive real interest rates because they can buy capital goods that provide a stream of returns more than sufficient to meet the interest cost.

Impatience to consume and the productivity of physical capital are the two forces that lead us to expect a positive real interest rate. Real interest rates are usually small and positive. Since real interest rates change little, big changes in nominal interest rates usually occur to offset big changes in inflation rates, keeping real interest rates in their normal range, determined by the forces of impatience and capital productivity. A good rule of thumb is that each percentage point rise in inflation is matched by a percentage point rise in nominal interest rates, leaving real interest rates the same as before.²

To calculate present values, we must be consistent. If we wish to calculate the present value of a future payment expressed in nominal terms, we must discount by the nominal interest rate. If the future payment is expressed in real terms, we must discount using the real interest rate.

The following is a common mistake. You want to buy a farm whose rental this year is £10 000. Today's interest rate is 10 per cent. You reckon that the farm's output should not change much over time. You use the formula of equation (1) for a perpetuity, divide £10 000 by 0.1, and get £100 000. The farmer wants £150 000 for the farm, so you decide not to buy.

You missed a financial killing. Nominal interest rates are 10 per cent only because the market thinks inflation will be about 7 per cent, leaving a real interest rate of 3 per cent. Doing the calculation in real terms at constant prices, we divide £10 000 for ever by 0.03 to obtain £333 000 as the right price for the farm. Equivalently, to calculate in nominal terms, we can use discount factors based on the 10 per cent nominal interest rate; however, remember that the likely inflation rate of around 7 per cent will steadily increase the nominal farm rental over time. If we do this calculation, we shall again conclude that £100 000 for the farm is a bargain.

11.3 Saving, investment and the real interest rate

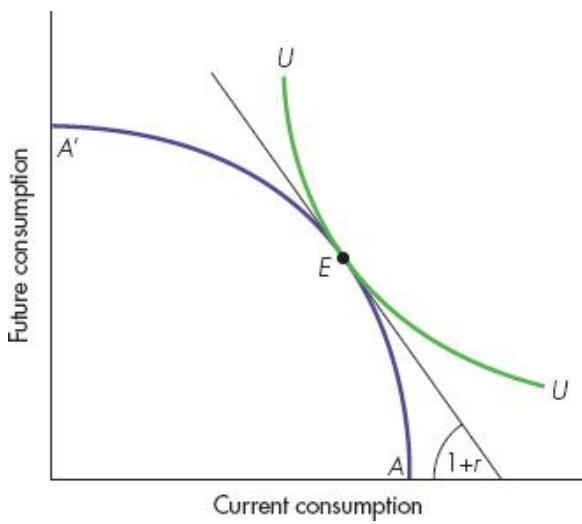
In this section we show how an economy decides how much capital stock to produce.

Figure 11.3 displays the production possibility frontier AA' , which shows the feasible combinations of current and future consumption goods that the economy can produce. The way agents can transfer consumption over time is through saving. At A the economy only produces for current consumption, at A' only for future consumption.

The frontier AA' shows different amounts of *investment* in the capital stock. At A not only is no investment undertaken, the existing capital is sold off and entirely consumed today. Future consumption is zero.

At A' all current resources are going in investment to raise the capacity to make consumption goods in the future. Current consumption is zero. Moving down AA' , more and more resources are transferred from future to current consumption. As usual, the curvature of the production possibility frontier reflects diminishing returns in this trade-off.

Assume for simplicity that inflation is zero, so that nominal and real interest rates coincide. The slope of the frontier is the extra future consumption from sacrificing a unit of current consumption. The slope has magnitude $-(1 + i)$, where i is the real rate of return on investment. The minus sign reminds us we sacrifice current consumption to add to future consumption.



By devoting more current resources to investment, society can trade off current for future consumption, moving up the frontier AA' . The frontier has slope $-(1 + i)$, where i is the rate of return on investment. Facing a real interest rate r , producers will choose E and so will consumers. The equilibrium real interest rate balances the productivity of investment and thriftiness of consumers.

Figure 11.3 The equilibrium real interest rate

What about consumer tastes? Both current and future consumption are desirable, so we can imagine a standard indifference map. The more impatient are consumers for current consumption, the steeper their indifference curves. Impatient people will give up lots of future consumption to get a bit more today. Thrifty people have flatter indifference curves.

In Figure 11.3, UU is the highest indifference curve that can be reached and E is the best allocation of current resources between consumption and investment.

The real interest rate may adjust to accomplish this outcome, even though decisions to add to the capital stock are taken by firms, and decisions about **saving** are taken by households.

Firms will invest until the real rate of return i equals the real interest rate r at which they can borrow money. Households face a budget line slope $-(1 + r)$, since by saving and lending they can exchange £1 of consumption today for $-(1 + r)$ of consumption in the future. Households save up to the point at which their indifference curve is tangent to their budget line with slope $-(1+r)$.

Saving is the difference between current income and current consumption.

Equilibrium occurs where saving equals investment. Households and firms are happy with the same transfer of resources from the present to the future. Figure 11.3 shows the equilibrium real interest rate r . Firms wish to be at E , where the rate

of return i equals the cost of borrowing r [the slope of the frontier $-(1+i)$ is tangent to the line $-(1+r)$]. Households want to be at the same point E, where their indifference curve UU is tangent to the line $-(1+r)$.

11.4 The demand for capital services

The analysis of the demand for capital services by an industry closely parallels the analysis of labour demand in Chapter 10. The rental rate for capital replaces the wage rate. Each is the cost of hiring factor services. We emphasize the *use* of *services* of capital. The example to bear in mind is a firm renting a vehicle or leasing office space. In demanding capital services, a firm considers how much extra output another unit of capital services will add.

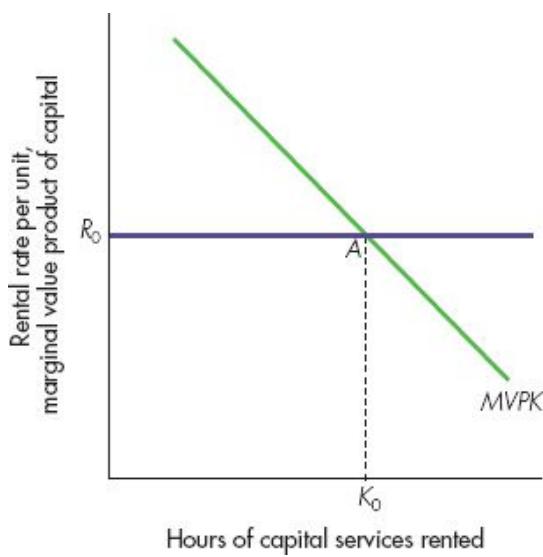
We can generalize our analysis to the case where the firm has monopoly power in its output market or monopsony power in its input market. Having discussed that extension in Chapter 10, we confine our discussion of capital services to the simpler case in which the firm is competitive.

Given the amounts of other inputs, the **marginal value product of capital (MVPK)** declines as more capital services are used. Although the firm's output price is fixed since it is competitive, the marginal physical product of capital is subject to diminishing returns. Figure 11.4 shows a downward-sloping MVPK curve, just like the MVPL curve in Chapter 10.

A firm rents capital up to the point at which its marginal cost – the rental rate – equals its marginal value product. The firm demands K_0 capital services at the rental rate R_0 .

The **marginal value product of capital (MVPK)** is the extra value of the firm's output when another unit of capital services is used, all other inputs being held fixed.

For given rental rates and quantities of other factors of production, *MVPK* is the firm's demand curve for capital services at each rental rate for capital services. The firm's *MVPK* curve showing its capital demand curve can be shifted outwards by one of three events: (1) an increase in its output price, which makes the marginal *physical* product of capital more valuable, (2) an increase in the level of other factors (chiefly labour) with which capital works to produce output, making capital more productive, or (3) a technical advance that makes capital more productive.



Diminishing marginal productivity implies a falling *MVPK* schedule as capital input is increased holding constant the quantity of other inputs. At any given rental, the firm hires capital services up to the point at which the rental per unit equals the *MVPK*. Thus the *MVPK* curve is also the firm's demand curve for capital services. For example, at a rental rate R_0 the firm will hire K_0 capital services.

Figure 11.4 The demand for capital services

The industry demand curve for capital services

As with labour, we can move from the firm's demand for capital services to the industry demand curve for capital services by horizontally adding the marginal value product of each firm. Again, we must recognize that, in expanding output, the industry bids down the price of its output.

Thus the industry demand curve for capital services is steeper than the horizontal sum of each firm's *MVPK* curves. The industry demand curve recognizes that output prices fall as output rises. The more inelastic the demand curve for the industry's output, the more inelastic is the industry's derived demand curve for capital services.

11.5 The supply of capital services

Capital services are produced by capital assets. We analyse the market for capital services, then consider what this implies for the market for capital assets. In so doing, we assume that the flow of capital services is directly determined by the stock of capital assets, such as machines.

This is a simplification. By working overtime shifts, a firm can alter the effective flow of machine services it gets from a given machine bolted to the factory floor. It can also leave machines idle.

Even so, in normal times firms have limited ability to vary the flow of capital services from a given capital stock. We shall grasp the key features of the market for capital if we assume that the flow of capital services is determined by the stock of capital available. Our analysis must distinguish between the long run and the short run, and examine both the supply of capital services to the economy and to a particular industry.

The short-run supply of capital services

In the short run, the total supply of capital assets (machines, buildings and vehicles), and thus the services they provide, is fixed to the economy. New factories cannot be built overnight. The supply curve for capital services is vertical at a quantity determined by the existing stock of capital assets. Some types of capital are fixed even for an individual industry. The steel industry cannot change overnight its number of blast furnaces. However, by offering a higher rental rate for delivery vans, the supermarket industry can attract a larger share of the delivery vans that the economy currently has. For such capital services, an industry faces an upward-sloping supply curve. It can bid services away from other industries.

The long-run supply of capital services

In the long run, the quantity of capital in the economy can be varied. New machines and factories can be built. Conversely, without new investment in capital goods the existing capital stock will depreciate and gradually fall. Similarly, individual industries can adjust their stocks of capital.

At what rental rates will owners of capital *assets* be willing to buy or build?

You buy a machine to rent out as a business. The machine costs £10 000, which you borrow. How much must the machine earn if you are to break even? First you have to cover the interest cost. Suppose the *real*, or inflation-adjusted, interest rate is 5 per cent. You have to pay the bank £500 ($= \text{£}10\,000 \times 0.05$) a year in real terms.

Then you have spending on maintenance. Also, the resale value of the machine depreciates each year. In real terms, maintenance and depreciation cost you £1000 per annum, 10 per cent of the purchase price. The depreciation rate is 10 per cent a year. The annual cost of renting out a machine is £500 for the opportunity cost of the funds and £1000 for depreciation.

To break even, the **required rental** is £1500 a year at constant prices. The asset cost £10 000. Hence the required real rate of return is 15 per cent a year.³ It is worth borrowing if the real interest rate on the loan is less than 15 per cent a year.⁴

The **required rental on capital** just covers the opportunity cost of owning the asset.

What determines the required rental?

The required rental rate, or cost of using capital, depends on three things: the price of the capital good, the real interest rate and the depreciation rate. Depreciation depends largely on technology; that is, on how fast the machine wears out with use and age. The real interest rate is determined by economy-wide forces and changes only slowly. Treating the depreciation rate and the real interest rate as given, we examine how the purchase price of capital goods affects the required rental on capital.

The long-run supply curve for the economy

In the long run, the quantity of capital services must earn the required rental. If it earns more, people will build extra capital goods. If it earns less, owners of capital will let assets depreciate without building new ones.

Figure 11.5 shows the long-run supply curve of capital services to the economy. Capital services come from capital goods. The construction industry produces buildings and the motor industry produces container lorries. Each industry has an upward-sloping supply curve. The higher the price of the capital good, the more the capital goods producing industry will choose to supply.

In the long run, a larger flow of capital services needs a higher capital stock. But capital depreciates. The higher the capital stock, the larger is total depreciation. Thus, a higher long-run flow of capital services needs a higher capital stock, which needs a higher flow of new capital goods to offset depreciation and maintain the capital stock intact.

Producers need a higher price for capital goods to make more new capital goods. To maintain the required rate of return on assets, we need a higher rental rate for capital services. In the long run a higher flow of capital services is supplied only if the rental rate on capital rises to match the higher price of capital goods needed to induce producers of new capital goods to keep pace with higher absolute levels of depreciation.

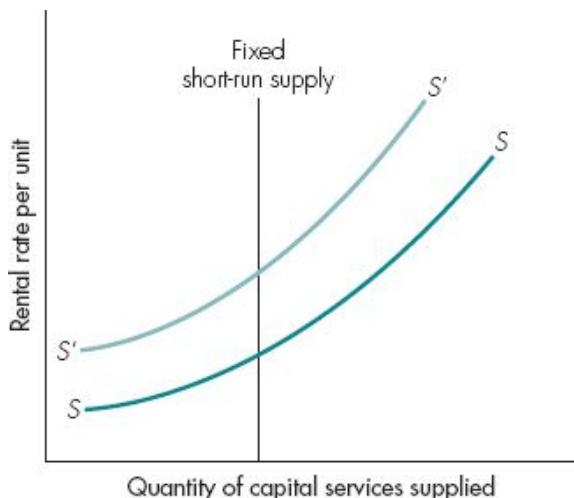
Figure 11.5 shows the long-run supply curve for capital services. SS slopes upwards in the long run when plotted against the rental rate on capital. We draw SS for a given real interest rate. If the real interest rate rises, the opportunity cost of holding capital assets rises. For a given purchase price of capital goods, the required rental must rise. Suppliers of capital services need a higher return to offset the higher opportunity cost of the money they tie up in purchasing capital goods.

In Figure 11.5 higher real interest rates shift leftwards along the long-run supply curve for capital services, from SS to $S'S'$. Higher rentals at each level of capital services (and hence capital assets) provide the higher real return to match the increase in the real interest rate.

The long-run supply curve for the industry

The preceding analysis determines the supply of capital services to the economy. In the long run, a small industry can get as much of this capital as it wishes, provided it pays the going rental rate. A larger industry may bid up the rental rate as it attracts a large fraction of the economy-wide supply of capital. Such an industry faces an upward-sloping supply curve for capital services.

We analyse the case of a small industry facing a horizontal long-run supply curve for capital services at the going rental rate. The analysis is easily extended to an industry facing an upward-sloping long-run supply curve for capital services.



In the short run, the stock of capital goods, and the services they supply, is fixed by past investment decisions; new capital goods cannot be produced overnight. In the long run, the higher rental rate is required to call forth a higher supply of capital services and a permanently higher capital stock. The higher rental rate just offsets the higher price for capital goods required to induce higher output of new capital goods to match the higher total depreciation of a larger capital stock. Thus the required rate of return is met at all points on SS . If real interest rates increase, the required rate of return will also increase to match the opportunity cost of funds tied up in capital goods. Hence the long-run supply curve of capital services shifts up to $S'S'$ providing a higher rental level at each level of the capital stock and its corresponding purchase price. Each point on $S'S'$ matches the new required rate of return.

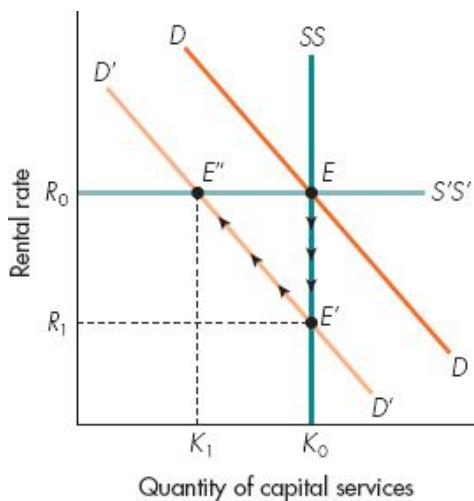
Figure 11.5 The supply of capital services to the economy

11.6 Equilibrium and adjustment in the market for capital services

Figure 11.6 shows the market for capital services for a particular industry. Long-run equilibrium is at E , where the horizontal long-run supply curve $S' S'$ crosses the industry demand curve DD derived from firms' $MVPK$ curves. The industry hires K_0 capital services at the going rental, R_0 .

Adjustments in the market for capital services

Suppose workers in the industry get a wage increase. In the long run this has a *substitution effect* and an *output effect*. The substitution effect makes firms switch to more capital-intensive techniques, raising the demand for capital services. However, by raising costs, a wage increase reduces the quantity of output supplied. This output effect reduces demand for all inputs. The second effect is more likely to dominate, the more elastic is the demand for the industry's output.



The industry begins in equilibrium at E . Overnight its short-run supply of capital is fixed at K_0 , but in the long run it faces the horizontal supply curve $S'S'$ at the going rental, R_0 . Suppose a wage increase shifts the demand curve for capital from DD to $D'D'$. The new short-run equilibrium is at E' . Since the rental R_1 fails to provide the required rate of return, owners of capital goods allow these goods slowly to depreciate without buying any new capital goods. The industry's capital stock and the services it provides gradually fall back. Eventually the industry reaches long-run equilibrium at E'' . Since capital is again earning the required rate of return, owners of capital goods now replace goods as they depreciate.

Figure 11.6 Short- and long-run adjustment of capital to a wage rise

Short-run and long-run adjustment

Suppose in Figure 11.6 the wage rise reduces demand for capital services from DD to $D'D'$. The industry begins in equilibrium at E . Initially, the short-run supply of capital services SS is vertical at K_0 . When demand shifts from DD to $D'D'$, the industry cannot immediately cut its input of capital services. With a vertical short-run supply curve, the new short-run equilibrium is at E' . The rental on capital falls from R_0 to R_1 .

The industry faces a long-run supply curve $S'S''$ for capital services. Eventually it must pay the going rate. At E' , owners of capital do not get the required rental for the capital services they supply. They let their capital stock depreciate. Over time, the industry's capital stock and supply of capital services fall until equilibrium is reached at E'' . The capital services used by the industry have fallen to K_1 . Less

capital means a higher marginal product of capital and higher rentals. At E'' users of capital again pay the required rental, R_0 .

The arrows in Figure 11.6 show the dynamic path that the industry will follow. When demand for capital falls, there is a sharp fall in the rental on capital. Owners of the fixed factor cannot adjust the quantity of capital services supplied. As time elapses, they adjust the quantity, allowing capital goods to depreciate, and the rental gradually recovers.

CONCEPT 11.1

FACTOR MARKETS: A SUMMARY

Chapter 10 and this chapter examine markets for production inputs. In the long run, when all inputs can be freely varied, the firm's choice of technique at each output level is determined by technology and relative factor rentals. At a given output, a higher relative price of one factor makes the firm substitute towards techniques using that factor less intensively. The long-run total cost curve shows the cheapest way to produce each output level when production techniques are optimally chosen.

From long-run total cost, we calculate long-run marginal cost and hence the output at which marginal cost and marginal revenue are equal. For each factor, the firm's demand is a derived demand that reflects the factor's marginal physical product in making extra output and the marginal revenue from selling that extra output. A competitive firm's demand curve for a factor is the marginal value product schedule, which assumes a given output price, given quantities of all other inputs, and given technology. Changes in any of these shift the marginal value product schedule. In the short run, a competitive firm demands that quantity of its variable factor which equates its marginal value product and its factor rental. In the long run, every factor can be varied. Each factor is demanded to the point at which its factor rental equals its marginal value product given the quantity of all other factors, *each having been adjusted in the same way*.

What distinguishes labour, capital and land is mainly the speed with which their supply can adjust. The input of casual labour on construction sites or during crop picking is easily variable, even in the short run. The supply of skilled workers with extensive training can be changed less quickly and the supply of capital goods takes even longer to adjust. Land is the factor whose total supply can never be adjusted. The slower the speed of adjustment, and the more irreversible the process, the more current decisions reflect beliefs about the future. The latter, neglected in our discussion of unskilled labour in Chapter 10, are central to our analysis of investment in physical capital.

11.7 The price of capital assets

We now turn from capital services to capital assets, demanded by firms wishing to supply capital services. Think of Hertz renting out cars, or property companies renting out office space. Anticipating a stream of rentals, suppliers of capital services work out the present value of this stream of rentals at the going interest rate. This tells us how much they should be prepared to pay to buy a capital asset. The price of capital assets is higher when (a) the anticipated rental stream is higher or (b) the interest rate is lower. Both raise the present value of the future rental stream.

People anticipating a higher stream of rental earnings pay a high purchase price for capital assets. At a lower price, people with lower anticipated streams then find it profitable to demand capital goods. There is a downward-sloping demand curve for capital goods. The lower the price, the higher the quantity demanded. The upward-sloping supply curve and downward-sloping demand curve together determine the equilibrium price and quantity of capital goods for the economy. This determines the flow supply of capital services that this stock will provide.

What happens when an individual industry faces a fall in its derived demand for capital services, as in Figure 11.6 ? In the short run, the rental on capital services falls to R_I . Moreover, everyone can work out that it will take some time before the rental rate climbs back to R_0 . At the going interest rate, the present value of rental earnings on new capital goods in this industry falls.

Now the industry is no longer willing to pay the economy-wide equilibrium price for capital assets. It makes no new investment, and its capital stock depreciates. Its capital stock falls until capital services become so scarce that the rental rate returns to its original level. The present value of future rentals then matches the price of capital goods in the whole economy. The industry now buys capital goods to replace goods as they depreciate. The capital stock is constant, and the industry is in its new long-run equilibrium.

The long-run equilibrium price of a capital asset is both the price that induces suppliers to make enough new assets to offset depreciation and keep the capital stock constant, and the price that buyers of capital goods are prepared to pay for that quantity. That price is the present value of the anticipated rental stream for capital services discounted at the going rate of interest.

11.8 Land and rents

Land is essentially a capital good in fixed supply to the economy, even in the long run. This is not literally true. The Dutch reclaimed from the sea some areas of low-lying land, and fertilizers enhance the effective input of land for farming.

Nevertheless, it makes sense to think about a factor whose total long-run supply is fixed.

CASE 11.1

THE BEST ADDRESS

Since land is in fixed supply, land prices are highest where demand is greatest. We should expect that the demand for land is higher in places like major cities where many people live. The table below shows the ten most expensive places to live according to the average prices of residential apartments in 2012. The price is for square metres and expressed in US dollars.

Source:

Top 10 most expensive cities	
1.Monaco (\$52.353)	6.Geneva (\$17.026)
2.London (\$25.789)	7.Singapore (\$16.350)
3.Hong Kong (\$20.371)	8.Tokyo (\$15.122)
4.Paris (\$19.843)	9.New York (\$13.377)
5.Moscow (\$17.566)	10.Mumbai (\$11.306)

source: www.globalpropertyguide.com.



A row of classic Edwardian houses in the area of Kensington and Chelsea, London. © David Palmer | istockphoto.com

Figure 11.7 shows the derived demand curve DD for land services. With a fixed supply SS , the equilibrium rental per acre is R_0 . A rise in the derived demand, for example because wheat prices rise, raises the rental to R_I . The quantity of land services is fixed by assumption.

Consider a tenant farmer who rents land. Wheat prices have risen but so have rents. Not only may the farmer be no better off, but the connection between the two rises may also be unrecognized. The farmer complains that high rents make it hard to earn a decent living. As in our discussion of footballers' wages in Chapter 10, it is the high derived demand combined with the inelastic factor supply that causes the high payments for factor services.⁵

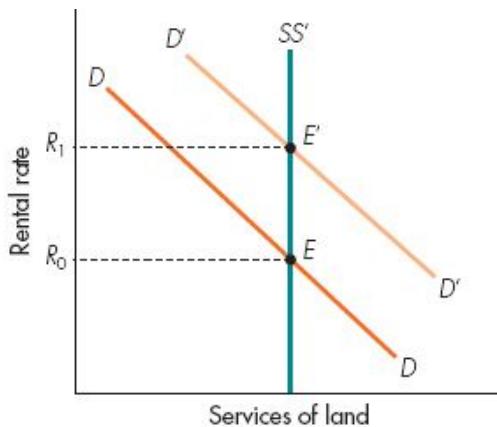


Figure 11.7 The market for land services

The total supply of land is fixed to the economy. The supply curve is vertical. The derived demand curve for land services reflects the marginal value product of land. Its derivation is exactly the same as the demand curves for labour and capital from the *MVPL* and *MVPK* schedules. The demand curve *DD* for land services determines the equilibrium land rental rate R_0 . If the derived demand curve for land services shifts up to *D'D'*, the equilibrium land rental will increase to R_1 .

Figure 11.7 The market for land services

Because land is *the* asset in fixed supply, economists have taken over the word 'rent', the payment for land services, to the concept of *economic rent*, the excess of actual payments over transfer earnings, introduced in Chapter 10. Economic rent is large when supply is inelastic.

11.9 Income distribution in the UK

The income of a factor is its rental rate multiplied by the quantity of the factor employed. We pull together our discussion of factor markets to examine the distribution of income in the UK.

The functional distribution of income

Table 11.6 shows the total earnings of the different factors of production in the UK during 2007–and compares their share of national income with the share they received during 1981–89. There have been small changes in the shares of different

factors of production income among different parts of those changes reflects the consequence of the financial crisis that hit the UK economy after 2007.

The **functional distribution of income** is the division of national income among different factors of production.

Table 11.6 UK functional income distribution, 1981–2012 (% of national income)

Source (factor of production)	1981–89 average	2007–12 average
Employment	64	62
Self-employment	7	13
Profits and property rents	29	25

Source: ONS, *UK National Accounts*.

Aggregate labour supply to the economy is relatively inelastic. The total number of employed workers was little higher in 2012 than in 1981. Table 11.2 shows that the UK capital per worker has increased. Technical progress has also boosted productivity. Total output increased and labour's marginal product schedule shifted outwards. Confronted with an almost vertical labour supply curve, this rise in the demand for labour raised the equilibrium real wage. When total output decreases, demand for labour decreases and so does the equilibrium real wage. Labour income from employment changes as national income changes.

Table 11.6 shows that the share of income from profits and rents fell slightly during 1989–2012, while income from self-employment increased. The quantity of capital employed rose steadily, at about the same rate as national output. Since the ratio of capital to output has been fairly constant, the real return on capital declines and so does the share of capital earnings in national output.

The supply of land is very inelastic. If national income increases, the derived demand curve for land shifts upwards. Property rentals increase in line with national income.

The personal income distribution

The **personal income distribution** is relevant to issues such as equality and poverty. Table 11.7 excludes the very poor, whose income is so low that HM Revenue and Customs does not record what they earn. Even confining attention to people who pay income tax, pre-tax income is unequal in the UK. Based on 30.5 million taxpayers, the top row of Table 11.7 shows that the poorest 900 000 households had an average taxable income of less than £7500 in 2010/11, whereas the bottom three rows show that the richest 2.8 million households all had taxable incomes in excess of £50 000.

The **personal income distribution** is the division of national income across individuals, regardless of the factor services from which these individuals earn their income.

Why do some people earn so much while others earn so little? Chapter 10 discussed some reasons why people earn different wages and salaries. Unskilled workers have little training and low productivity.

Table 11.7 UK personal income distribution, 2010/11

Taxable income band (£000 per annum)	Million taxpayers
< 7.5	0.9
7.5–10	2.6
10–15	6.4
15–20	5.2
20–30	6.9
30–50	5.6
50–100	2.1
100–200	0.5
200+	0.2

Source: ONS, *Social Trends*.

Workers with high levels of training and education earn much more. Some jobs, such as coal mining, pay high compensating differentials to offset unpleasant working conditions. Pleasant, but unskilled, jobs pay much less since many people are prepared to do them. Talented superstars in scarce supply but strong demand earn very high economic rents.

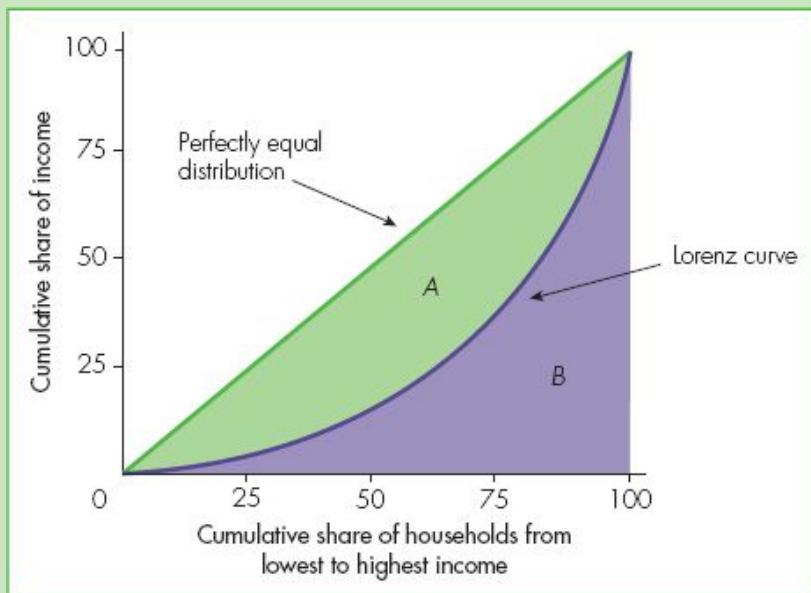
CONCEPT 11.2

HOW TO MEASURE INCOME INEQUALITY: THE GINI COEFFICIENT

Is income in the UK more equally distributed than in Brazil? Is income in the US more unequally distributed today than it was in the past?

To answer those questions we need a statistical measure of income inequality. This measure is provided by the Gini coefficient, proposed by the Italian statistician Corrado Gini in 1912. The Gini coefficient measures how an observed income distribution in a given country is different from a perfectly equal distribution (that is, a distribution where income is divided equally among all individuals). Suppose we observe a given distribution of income. A perfectly equal distribution would imply that each 1 per cent of the population

has 1 per cent of total income. This is shown as the straight line in the figure below. That line is a 45-degree line. The graph of the actual observed income distribution is called the Lorenz curve. Points on the Lorenz curve represent statements such as ‘the bottom 10 per cent of all households earn 10 per cent of total income’ or ‘the top 50 per cent of all households earn 70 per cent of total income’, and so on. The larger is the difference between the Lorenz curve and the 45-degree line, the more unequally is income distributed. This difference is given by area *A* in the figure.



The Gini coefficient is calculated as:

$$Gini = \frac{Area|A|}{Area|A| + Area|B|}$$

In the case of a perfectly equal distribution, the Lorenz curve coincides with the 45-degree line and so area *A* is zero. The Gini coefficient in this case is zero. In the case of perfectly unequal distribution (that is, a single individual has all the income and everybody else has zero income), area *B* will be zero and so the Gini coefficient is 1. Thus a Gini coefficient closer to 1 means higher income inequality.

The Gini coefficients for selected countries in the late 2000s are displayed below.

Country	Gini coefficient
Brazil	0.52
Denmark	0.25
UK	0.34
US	0.38

Source: OECD; CIA, *World Factbook*.

According to the data, income is more unequally distributed in Brazil than in Denmark. The UK and the US have similar income inequality.

Table 11.7 refers not just to income from the supply of labour services. One reason why the distribution of personal income is so unequal is that the ownership of wealth, which provides income from profits and rents, is even more unequal. Table 11.8 gives details for 2010.

Table 11.8 UK distribution of total wealth, 2010

Percentage of wealth owned by:	2010
Most wealthy 10%	44
Most wealthy 20%	62
Most wealthy 50%	90
Total wealth (£bn)	10 329.5

Source: ONS, *Wealth and Asset Survey*

The most wealthy 10 per cent of the population owned 44 per cent of UK total wealth in 2010 and the most wealthy 50 per cent of the population owned 90 per cent of UK total wealth. The stream of profit and rent income to which such wealth gives rise plays a large part in determining the personal distribution of pre-tax *income*.

Summary

- **Physical capital** comprises real assets yielding services to producing firms or consuming households. The main categories of physical capital are plant and machinery, residential structures, other buildings, consumer durables and inventories. **Tangible wealth** is physical capital plus land.
- **Present values** convert future receipts or payments into current values. Because lenders can earn – and borrowers must pay – interest over time, a pound tomorrow is worth less than a pound today. How much less depends on the interest rate. The higher the interest rate, the lower the present value of any future payment.
- Since lending or borrowing cumulates at compound interest, for any given annual interest rate the present value of a given sum is smaller the further into the future that sum is earned or paid.

- The present value of a **perpetuity** is the constant annual payment divided by the rate of interest (expressed as a decimal fraction).
- **Nominal interest rates** measure the monetary interest payments on a loan. The inflation-adjusted **real interest rate** measures the extra goods a lender can buy by lending for a year and delaying purchases of goods. The real rate of interest is the nominal interest rate minus the inflation rate over the same period.
- In the long run, the real interest rate adjusts to make investment equal to saving, and is determined by the return on firms' investments and the degree of impatience of households.
- The demand for capital services is a derived demand. The **firm's demand for capital services** is its marginal value product of capital curve. Higher levels of the other factors of production and higher output prices shift the derived demand curve up. The **industry demand for capital services** is less elastic than the horizontal sum of each firm's curve because it also allows for the effect of an industry expansion in bidding down the output price.
- In the short run, the supply of capital services is fixed. In the long run, it can be adjusted by producing new capital goods or allowing the existing capital stock to depreciate.
- The **required rental** is the rental that allows a supplier of capital services to break even on the decision to purchase the capital asset. The required rental is higher, the higher is the interest rate, the depreciation rate or the purchase price of the capital good.
- A rise in the industry wage has two effects on the derived demand curve for capital services. By reducing labour input, it reduces the marginal physical product of capital. By reducing industry output, it increases the output price. When output demand is very inelastic, the latter effect will dominate. When output demand is very elastic, the former effect dominates.
- The **asset price** is the price at which a capital good is bought and sold outright. In long-run equilibrium, it is both the price at which suppliers of capital goods are willing to produce and the price at which buyers are willing to purchase. The latter is merely the present value of anticipated future rentals earned from the capital services that the good provides in the future.
- **Land** is the special capital good whose supply is fixed even in the long run. However, land and capital can move between industries in the long run until rentals on land or on capital are equalized in different industries.

- Technology and the ease of factor substitution dictate the very different capital intensity of different industries. Most industries are becoming more capital intensive over time, but at different rates. This reflects the ease with which industries can substitute capital for labour, the rise in wage rates relative to capital rentals, and technical advances in different industries.
- The **functional distribution of income** shows how national income is divided between the factors of production. The share of each factor has remained fairly constant over time. This conceals a rise in the quantity of capital relative to labour, and a corresponding fall in the ratio of capital rentals to labour wages.
- The **personal distribution of income** shows how national income is divided between different individuals regardless of the factor services from which income is earned. A major cause of income inequality in the UK is a very unequal distribution of income-earning wealth.

Review questions



EASY

- 1 (a) Consumer durables, such as washing machines, are part of the capital stock but do not generate any financial income for their owners. Why do we include consumer durables in the capital stock? (b) To wash your clothes you can take them to a launderette and spend £2 per week indefinitely or buy a washing machine for £400. It costs £1 per week (including depreciation) to run a washing machine, and the interest rate is 10 per cent per annum. Does it make sense to buy the washing machine? Does this help you answer part (a)?
- 2 A bank offers you £1.10 next year for every £0.90 you give it today. What is the implicit interest rate?
- 3 The interest rate falls from 10 to 5 per cent. Discuss in detail how this affects the rental on capital services and the level of the capital stock in an industry in the short and long run.
- 4 Suppose a plot of land is suitable only for agriculture. Can the farming industry experience financial distress if there is an increase in the price of land? Is your answer affected if the land can also be used for housing?
- 5 **Common fallacies** Why are these statements wrong? (a) Inflation leads to high nominal interest rates. This reduces the present value of future income. (b) If the economy continues to become more capital intensive, eventually there will be no jobs left for workers to do. (c) Since the economy's supply of land is fixed, it would be supplied even at a zero rental, which should therefore be the equilibrium rental in the long run.

MEDIUM

6 Suppose you face the following two different investment opportunities:

- (i) You can invest £3000 and after five years you are going to get £4500.
- (ii) You can invest £3000 at the annual market interest rate of 10 per cent for five years. Which investment will you opt for? Demonstrate your answer by comparing the present values of the two investment opportunities.

7 Suppose that the *real* interest rate in the economy is 4 per cent, while the inflation rate one year from now is known to be 2 per cent. Use the Fisher equation to find the nominal interest rate. Use the nominal interest rate to find the present value of £100 one year from now. Now suppose that inflation in one year from now is known to be 4 per cent. How has the present value calculated previously changed? Why?

8 A firm buys a machine for £10 000, earns rentals of £3600 for each of the next two years and then sells it for scrap for £9000. Use the data in the table below to determine if the machine is worth buying when the interest rate is 10 per cent per annum.

Interest and present value (PV)

	Year		
	0	1	2
<i>At 10% interest rate:</i>			
Value of £1 lent today in:	£1	£1.10	£1.21
PV of £1 earned in:	£1	£0.91	£0.83
<i>At 5% interest rate:</i>			
Value of £1 lent today in:	£1	£1.05	£1.10
PV of £1 earned in:	£1	£0.95	£0.91

MEDIUM

9 Suppose you buy a machine that costs £10 000 today and rent it out. You earn a rent of £1500 on that machine every year for four years. After four years the machine can be sold as scrap for £4000. Assume that the interest rate is 10 per cent in all four years. What is the present value of the machine? What is the net present value of the machine? Is the investment worthwhile?

10 Suppose that the demand for capital is given by $K = 20 - 2r$, where K is capital and r is the rental rate. In a graph with r on the vertical axis and K on the horizontal axis, plot the demand for capital. Suppose that in the short run the supply of capital is fixed at 6 units. In a graph show how the rental rate is determined in equilibrium. An earthquake destroys part of the capital available in the economy. The supply of capital shrinks to 4 units in the short run. What happens to the equilibrium rental rate?

|1 What should be the impact of globalization on assets in fixed supply, particularly land? Can you think of an example in which globalization might induce a fall in land prices?

HARD

- |2 Show on a graph how society's trade-off between saving and consumption helps a firm make a decision on much to invest in capital stock.
- |3 A firm is producing output using only capital. Its production function is $Q = 10K - K^2$. The firm sells its product in a competitive market at a price of £2 and it rents capital from a competitive market at a rental rate of r per unit of capital. Write down the profit function of the firm and find its capital demand function
- |4 Suppose you buy a car and decide to rent it out for a fee. The car costs you £5000. You can buy the car only by borrowing from the bank. How would you decide if you are better off renting or owning the car?

-
- 1 From Irving Fisher, the economist who first pointed out that relationship between nominal and real interest rates.
 - 2 While this is the usual case, under particular economic circumstances real interest rates can become negative. An example is given by the recession resulting from the credit crunch. Short-term nominal interest rates were pushed to almost zero after 2007, as you can see in Figure 11.2. With positive inflation and a zero nominal interest rate, the real interest rate can be negative.
 - 3 To simplify the calculation, assume the machine and the bank loan last for ever. We can then use our formula for the present value of a perpetuity. The price p of a perpetuity is the annual payment c divided by the required rate of return r that lenders could get by lending to a bank. If $p = c/r$; then $r = c/p$. When c is the annual cost and p is the initial price of a machine, you need a rate of return $r = c/p$ to make it worth renting out machines.
 - 4 If the firm using the capital services also owns the capital good, the required rental is the cost the firm should charge itself to use the capital when calculating economic costs (see our discussion of accounting versus economic costs in Chapter 6).
 - 5 If most farmers rent their land, agricultural subsidies, such as the EU Common Agricultural Policy, increase land rentals as well as the price farmers get from crops. It is the landowners who really benefit!

CHAPTER 12

Risk and information

Learning outcomes

By the end of this chapter, you should be able to:

- 1 describe risk aversion and diminishing marginal utility
- 2 define risk pooling and risk spreading
- 3 understand how inside information leads to moral hazard and adverse selection
- 4 analyse how an asset return reflects its cash income and its capital gain (loss)
- 5 describe how correlation of asset returns affects risk pooling
- 6 explain asset market efficiency
- 7 recognize spot and forward markets

Every action today has a future outcome that is not certain. It is risky. When you start studying economics, you have only a rough idea of what is involved, and even less idea about how it will be used once the skill is acquired. This chapter examines how risk affects our actions, and how economic institutions have evolved to help us deal with the risky environment in which we live. The importance of risk in economics has been magnified by the recent credit crunch. A global economic crisis was created, among other factors, by the excessive risks taken by large international investors.

Some activities reduce risk, but others increase it. We spend money on insurance, but also on the lottery and on risky assets in the stock market. People generally dislike risk and are prepared to pay to have their risks reduced. This explains the existence of many economic institutions that, at a price, allow people who dislike risk most to pass on their risks to others more willing or more able to bear them.

12.1 Individual attitudes to risk

A risky activity has two characteristics: the likely outcome and the degree of variation in the possible outcomes. Suppose you are offered a 50 per cent chance of

making £100 and a 50 per cent chance of losing £100. On average, you make no money by taking such gambles.

A **fair gamble** on average yields zero monetary profit.

A **risk-neutral** person is interested only in whether the odds yield a profit on average.

In contrast, a 30 per cent chance of making £100 and a 70 per cent chance of losing £100 is an *unfair* gamble. On average, you lose money. With the probabilities of winning and losing reversed, the gamble would on average be profitable. The odds are then *favourable*.

Compare a gamble with a 50 per cent chance of making or losing £100 and a gamble with the same chances of winning or losing £500. Both are **fair gambles**, but the second is *riskier*. The range of possible outcomes is greater.

We turn now to individual tastes. Economists classify people as risk-averse, risk-neutral or risk-loving. The key issue is whether or not a person would accept a fair gamble. A **risk-neutral** person ignores the dispersion of possible outcomes, betting if and only if the odds on a monetary profit are favourable.

A **risk-averse** person may bet if the odds are very favourable. The probable monetary profit overcomes the inherent dislike of risk. The more risk-averse the individual, the more favourable must be the odds before she takes the bet.

The more **risk-loving** the individual, the more unfavourable must be the odds before the individual will not bet.

A **risk-averse** person will refuse a fair gamble.

A **risk-lover** bets even when the odds are unfavourable.

Insurance is the opposite of gambling. Suppose you own a £100 000 house. There is a 10 per cent chance it burns down by accident. You have a 90 per cent chance of continuing to have £100 000 but a 10 per cent chance of having nothing. Our risky world forces you to take this bet. On average, you end up with £90 000, which is 90 per cent of £100 000 plus 10 per cent of nothing.

An insurance company offers to insure the full value of your house for a premium of £15 000. Whether or not your house burns down, you pay the insurance company the £15 000 premium. It pays you £100 000 if it burns down. Whatever happens, you will end up with £85 000.

Would you insure? The insurance company is offering unfavourable odds, which is how it makes its money. Uninsured, on average you are worth £90 000, insured only £85 000. A risk-neutral person would not insure on these terms. The mathematical calculation in monetary terms says it is on average better to stand the risk of a fire. The risk-lover will also decline insurance. Not only are the odds poor,

there is also the added enjoyment of standing the risk. But a person who is sufficiently risk-averse will accept the offer, happy to give up £5000 on average to avoid the possibility of catastrophe. Table 12.1 summarizes this discussion of attitudes to risk.

Table 12.1 Attitudes towards risk

Tastes	Betting	Insurance at unfair premium
Risk-averse	Needs favourable odds	May buy
Risk-neutral	Except at unfavourable odds	Won't buy
Risk-lover	Even if odds against	Won't buy

Diminishing marginal utility

People's tastes exhibit a **diminishing marginal utility of income**. Successive equal rises in income quantities add less and less to total utility.

An important aspect of understanding how people behave when facing risk is the concept of **diminishing marginal utility of income**. In Chapter 5 we discuss the idea of decreasing marginal utility for consumption goods. Here, we see how the same principle can be applied to individuals' income when those individuals face risky choices.

Suppose you are starving and broke. Getting £1000 would yield you a lot of utility or happiness, by allowing you to buy basic goods you really need. If you then got another £1000, there are still things to spend it on. However this extra £1000, while still helpful, brings you less extra value than when you were desperate.

Thus, the marginal utility of the first £1000 is very high. You really needed it. The marginal utility of the next £1000 is not quite so high. As you get more, the marginal utility of the extra consumption tends to diminish.

Of course, there are exceptions to this general rule. Some people *really* want a yacht, and their utility takes a huge jump when they can finally afford one. But most of us first spend our money on things we most need, and get less and less extra satisfaction out of successive equal increases in our spending power.

You have £11 000 and are offered an equal chance of winning or losing £10 000. This is a fair bet in money terms since the average profit (or expected income) is £0. But it is not a fair bet in utility terms. Diminishing marginal utility implies that the extra utility you enjoy if the bet wins, taking your total wealth from £11 000 to £21 000, is much smaller than the utility you sacrifice if the bet loses, taking your wealth from £11 000 to £1000. You get a few extra luxuries with the £10 000 you

might win, but you have to give up almost everything if you lose and have to survive on only £1000.

A risk-averse person declines a fair bet in money terms. The hypothesis of diminishing marginal utility implies that, except for the occasional gamble for pure entertainment, people should generally be risk-averse. They should refuse fair money gambles because they are not fair utility gambles. As we shall see, this story fits many of the facts.

Two implications of this analysis recur throughout the chapter. First, *risk-averse people devote resources to finding ways to reduce risk*. As the booming insurance industry confirms, people will pay to get out of some of the risks that the environment otherwise forces them to bear. Second, *individuals who take over the risk have to be rewarded for doing so*. Many economic activities consist of the more risk-averse bribing the less risk-averse to take over the risk.

CASE 12.2

WHY PLAY A LOSING GAME ? THE CASE OF THE NATIONAL LOTTERY

The UK's original National Lottery game, first introduced in 1994, is based on drawing six balls without replacement from a stock of 49 balls. The odds on matching all six balls are about 1 in 14 million. Only 45 per cent of sales revenue is returned as prizes, worse odds than received by a blind punter at a horse race. Nevertheless, the National Lottery is very popular and represents an important source of revenue for the government. Moreover, the poorest 20 per cent of the population in the UK account for over a third of all spending on the National Lottery.

Why do many low-income people buy lottery tickets even though the expected return from them is so low ? Recent research has shown that low-income people see playing the lottery as their best opportunity for improving their financial situation, albeit this is not rational. The hope of getting out of poverty encourages people to continue to buy tickets, even though their chances of stumbling upon a life-changing windfall are nearly impossibly slim and buying lottery tickets in fact exacerbates the very poverty that purchasers are hoping to escape. In practice, lottery tickets are an *inferior good*.

According to this research, lotteries set off a vicious cycle that not only exploits low-income individuals' desire to escape poverty but also directly prevents them from improving upon their financial situation. The government is thus concerned with the need to explore strategies that balance the economic burdens faced by low-income households with the need to maintain important funding streams.



© Sean Gladwell | Dreamstime.com

Source: E. Haisley et al., 'Subjective relative income and lottery ticket purchases', *Journal of Behavioral Decision Making* 21, no. 3 (2008): 283–295.

12.2 Insurance and risk

A farmer and an actress have risky incomes. Each gets 10 in a good month, 0 in a bad month. But the risks are *independent*. Whether the farmer has a good month is not connected to whether the actress has a good month. Individually, their incomes are very risky. Collectively, they are less so.

In Table 12.2 they *pool* their incomes and their risk, each getting half of their joint income. If they both have a good month (top-left entry) or both have a bad month (bottom-right entry), the pooling arrangement makes no difference. They each get what they would have got on their own. In the other two cases, the success of one partner offsets the failure of the other. Together, their income is more *stable* than as individuals. If the farmer and the actress are risk-averse, they gain by pooling their risky incomes. If it were not so hard to set up such deals (lawyers' fees, the problem of cheating, tax problems), we would see more of them.

Table 12.2 Sharing joint incomes by risk pooling

		Farmer
Actress		Good month, 10
Good month, 10		Bad month, 0
Good month, 10	10	5
Bad month, 0	5	0

Pooling independent risks is the key to insurance. Suppose mortality tables show that on average 1 per cent of people aged 55 will die during the next year. Deaths result from heart disease, cancer, road accidents and other causes, in predictable proportions.

Now randomly choose any 100 people aged 55 knowing nothing about their health. Throughout the nation, 1 per cent of such people will die in the next year. In our sample of 100 people, it could be 0, 1 or 2 per cent, or even more. The larger the

sample, the more likely it is that around 1 per cent will die in the next year. With 1 million 55 year olds we could be pretty confident that around 10 000 will die, though we could not of course say which ones. By putting together more and more people, we reduce the risk or dispersion of the aggregate outcome.¹

Life assurance companies take in premium payments in exchange for a promise to pay a large amount to the family if the insured person dies. The company can make this promise with great certainty because it pools risks over many clients. Since the company cannot guarantee that exactly 1 per cent of its many 55 year olds will die in any one year, there is a small element of residual risk for the company to bear, and it makes a small charge for this in calculating its premiums. However, the company's ability to pool the risk means that it will make only a small charge. If life assurance companies try to charge more, new entrants join the industry knowing that the profits more than compensate for the small residual risk to be borne.

Risk pooling aggregates independent risks to make the aggregate more certain.

Risk pooling does not work when all individuals face the same risk. Suppose there is a 10 per cent risk of a nuclear war in Europe in the next ten years. If it happens, everyone in Europe dies, leaving money to their nearest surviving relatives in the rest of the world. Ten million people in Western Europe offer to buy insurance from an American company.

Despite the number of people, the risk cannot be pooled. If everybody in Europe dies, if anybody dies, the insurance company either pays out to everybody's relatives or it pays nothing. In the aggregate, there is still a 10 per cent chance of having to pay out, just as individual Europeans face a 10 per cent chance of disaster. When the same thing happens to everybody, if it happens at all, the aggregate behaves like the individual. There is no risk reduction from pooling.

Many insurance companies do not insure against what they call 'acts of God' – floods, earthquakes, epidemics. Such disasters are no more natural or unnatural than a heart attack. But they affect large numbers of the insurance companies' clients if they happen at all. The risk cannot be reduced by pooling. Companies cannot quote the low premium rates that apply for heart attacks, where risks are independent and the aggregate outcome is fairly certain.

Risk sharing works by reducing the stake.

There is another way to reduce the cost of risk bearing. This is known as **risk sharing**, and the most famous example is the Lloyd's insurance market in London.

Risk sharing is necessary when it has proved impossible to reduce the risk by pooling. Lloyd's offers insurance on earthquakes in California, and insurance of a film star's legs.

To understand risk sharing, we return to diminishing marginal utility. We argued that the utility benefit from an extra £10 000 is less than the utility sacrificed when £10 000 is given up. However, this difference in marginal utility for equivalent monetary gains and losses is tiny if the size of the stake is tiny. The marginal utility from an extra £1 is only fractionally less than the utility lost by sacrificing £1. For small stakes, people are almost risk-neutral. You would probably toss a coin with us to win or lose £0.10, but not to win or lose £10 000. The larger the stake, the more diminishing marginal utility bites.

You go to Lloyd's to insure the US space shuttle launch for £20 billion – a big risk. Only part of this risk can be pooled as part of a larger portfolio of risks. It is too big for anyone to take on at a reasonable premium.

The Lloyd's market in London has hundreds of 'syndicates', each a group of 20 or so individuals who have each put up £100 000. Each syndicate takes perhaps 1 per cent of the £20 billion deal and then resells the risk to yet other people in the insurance industry. By the time the deal has been subdivided and subdivided again, each syndicate or insurance company holds a tiny share of the total. And each syndicate risk is further subdivided among its 20 members. The risk is shared out until each individual's stake is so small that there is a tiny difference between the marginal utility from a gain and the marginal loss of utility in the event of a disaster. It now takes only a small premium to cover this risk. The package can be sold to the client at a premium low enough to attract the business.

By pooling and sharing risks, insurance allows individuals to deal with many risks at affordable premiums. But two things inhibit the operation of insurance markets, reducing the extent to which individuals can use insurance to buy their way out of risky situations.

MATHS 12.1

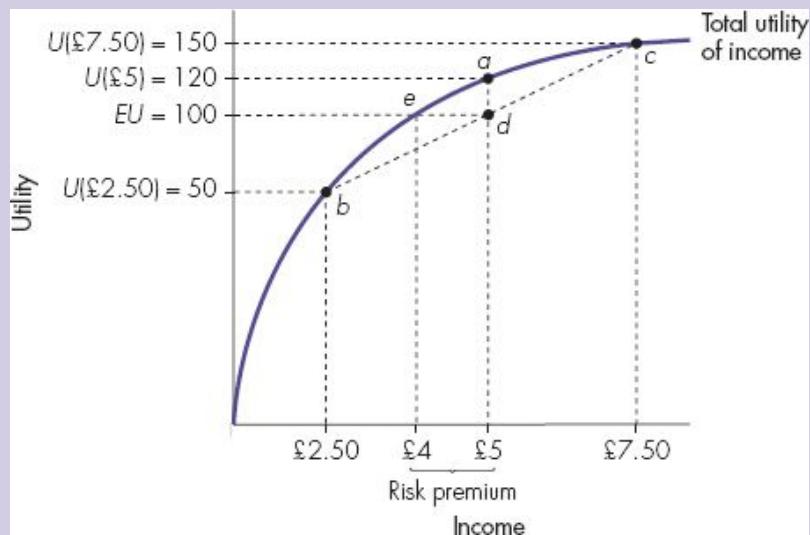
CHOICE UNDER UNCERTAINTY: EXPECTED UTILITY OF INCOME AND ATTITUDE TO RISK²

We can describe an individual's attitude to risk using the concept of total utility of income. This is a way to link the idea developed in Chapter 5 to the case of uncertainty.

Suppose that a consumer has an initial income of £5. He can invest his initial income in a stock asset today (assume that the stock today costs exactly £5). Next month the value of the stock can either increase or decrease. If it increases, then the consumer increases his income to £7.50. If the value of the stock decreases, the consumer ends up with an income of £2.50. Our consumer believes that there is a 50 per cent chance that the stock will increase its value in the next month. The expected value from buying the

stock asset is the probability weighted average of the value from each possible outcome: $EV = 0.5(\text{£}7.50) + 0.5(\text{£}2.50) = \text{£}5$. Buying the asset is, for our consumer, a *fair gamble*; on average, he does not gain or lose. Should the consumer buy the stock asset?

The answer depends on our consumer's attitude to risk. To understand preferences in relation to risk, we introduce the idea of utility of income for our consumer.



The utility of having £5 with certainty (point a) is higher than the expected utility from buying the asset (point d). The consumer is risk averse. Point d lies on the curve connecting points b and c. This is because expected utility is given by: $EU = 0.5U(\text{£}2.50) + 0.5U(\text{£}7.50)$.

The consumer obtains utility from income.³ Suppose that his utility of income is such that he obtains 120 in terms of utility if his income is £5, while he gets a utility of 150 if his income is £7.50 and a utility of 50 if his income is £2.50. More formally, we can write his utility as a function of income as: $U(\text{£}5) = 120$, $U(\text{£}7.50) = 150$ and $U(\text{£}2.50) = 50$. What is the expected utility level for the uncertain event faced by the consumer? The *expected utility* is the probability weighted average of the utility from each possible outcome. The expected utility of buying the stock is:

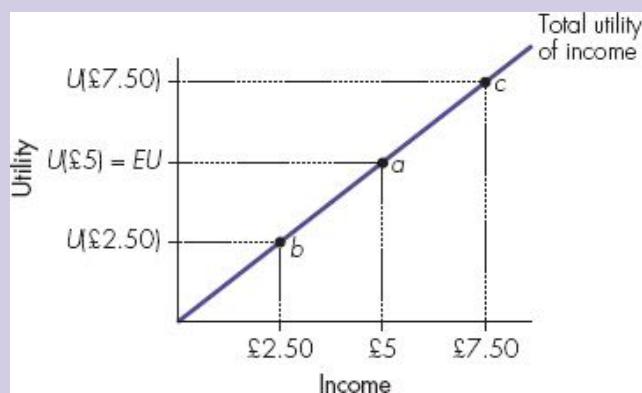
$$EU = 0.5[U(\text{£}7.50)] + 0.5[U(\text{£}2.50)] = 0.5(150) + 0.5(50) = 100$$

where EU denotes expected utility. For our consumer, the expected utility from buying the asset (100) is lower than the utility of not buying the asset and keeping £5 in his pocket with certainty (120). This implies that our consumer *prefers* having £5 with certainty (not buying the stock asset) to buying the asset, even if the asset gives him the same income on average. He does not like to take the risk. In this case, we say that our consumer is *risk-averse*. If we plot the total utility of income for a risk-averse consumer, the graph looks like the one in the following figure.

The utility from having £5 with certainty is point *a*, which is above point *d*, which represents the expected utility from buying the asset. Our consumer will not buy the asset in this case. The utility of income displayed in the figure has the property of diminishing marginal utility of income. As income increases, total utility increases by less and less.

A risk averse consumer is willing to pay to avoid bearing risk. The amount he would like pay is called the *risk premium*. Suppose that our consumer gets a utility of 100 if his income is £4; that is, $U(\text{£}4) = 100$. This is point *e* in the above figure. Now the expected utility from buying the asset is the same as the utility of having £4. In this case, our consumer is indifferent between taking the risk (that is, buying the asset) and having £4 with certainty. To avoid taking the risk, our consumer is willing to pay a risk premium of £1 = $\text{£}4 - \text{£}3$. If insurance is available, a risk-averse consumer will insure himself against risk.

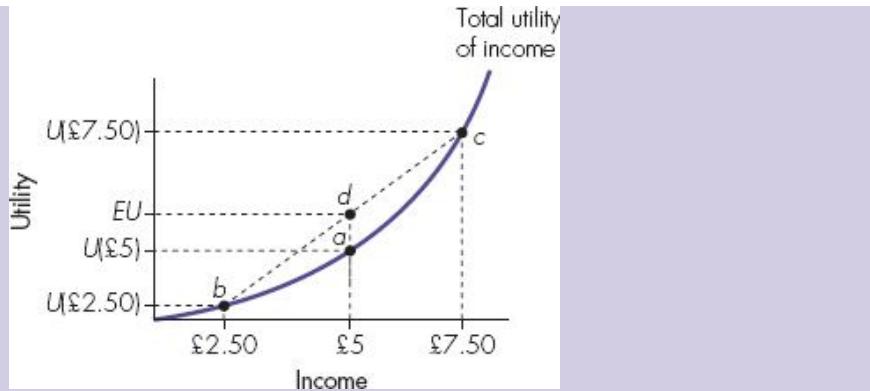
A consumer who is *risk-neutral* will have a total utility of income that looks like a straight line, as shown in the figure below.



The utility of having £5 with certainty (point *a*) is exactly equal to the expected utility from buying the asset. The consumer is risk-neutral.

In the case of a consumer who is risk-neutral, the utility of having £5 with certainty is equal to the expected utility of buying the asset. Here, the consumer is indifferent between buying and not buying the asset. In the case of a risk-neutral consumer, his total utility of income displays constant marginal utility of income.

Finally, in the following figure we plot the utility of income for a *risk-loving* consumer. A risk-lover prefers the fair gamble (in this case, buying the asset).



The utility of having £5 with certainty (point a) is lower than the expected utility from buying the asset (point d). The consumer is a risk-lover.

12.3 Asymmetric information

A concept that is related to risk and uncertainty is asymmetric information.⁴ Suppose you have a company and you want to hire a new worker. You would like to hire someone who is productive since you do not want to pay for a lazy worker. However, you do not know with certainty whether the worker you are hiring is going to be productive or not. On the other hand, the worker knows himself and knows his own productivity. In this case, there is asymmetric information, since one party (the worker) has more information about an important characteristic of the transaction (his productivity) than the other party (you, as the owner of the company). Why is this related to risk and uncertainty ?

You are uncertain about the productivity of the worker and, in the case of them being unproductive, you face the risk of paying someone who does not contribute to the profits of the company. The existence of asymmetric information creates the possibility of *opportunistic behaviour*: the informed individual benefits at the expense of the less-informed individual.

In the case of asymmetric information, we distinguish between two possible cases:

1. *Moral hazard (or hidden action)*: In this case, the uninformed agent cannot observe a particular *action* of the informed individual. For example, a worker may put little effort into performing his job if it is difficult for the employer to monitor him. The problem of moral hazard is also known as the *principal–agent problem*, where the principal is the name we give to the uninformed individual and the agent is the informed individual.
2. *Adverse selection (or hidden information)*: In this case, the uninformed individual does not know about an *unobservable characteristic* of the informed individual. For example, a person who wants to buy life insurance has more information about his own health than does the insurance company.

In the following sections we discuss the two cases in the context of the insurance market.

Moral hazard

Insurance companies calculate the statistical chances of particular events. They work out how many cars are stolen each year. Since different thefts are largely independent risks, we expect insurance firms to pool the risk over many clients and charge low premiums for car theft.

Sitting in a restaurant, you remember that your car is unlocked. Do you abandon your nice meal and rush outside to lock it? Not if you know the car is *fully* insured against theft. If the act of insuring changes the odds, then we have the problem of *moral hazard*. The informed individual (the insured) has an incentive to engage in risky behaviour; in this case, by not making the effort to minimize the chance of having her car stolen.

Statistical averages for the whole population, some of whom are uninsured and take greater care, are no longer a reliable guide to the risks the insurance company faces and the premiums it should charge. Moral hazard makes it harder to get insurance and more expensive when you do get it.

Insurance companies insure your car or house only up to a certain percentage of its replacement cost. They take over a big part of the risk, but you are worse off if the bad thing happens. The company gives you an incentive to minimize the chance of the bad thing happening. By limiting moral hazard, the insurance company pays out less frequently and can charge a lower premium.

CASE 12.2

WHY ARE CEO_S PAID SO MUCH ?

The compensation received by chief executive officers (CEOs) of large banks has featured prominently in the news over the past few years. As a result of the credit crunch, some troubled large banks were rescued using taxpayers' money. Nevertheless, those large banks continued to provide large bonuses to their CEOs. US President Barak Obama once described such large bonuses as 'shameful' and the public appears outraged each time one is announced in the news.

Executive compensation is a classic moral hazard or principal–agent problem. The principals (the shareholders) need to structure rewards for agents (managers) so that they make decisions that are in the long-term interests of the principals when the strategies of the agents are hardly discernible.

Lots of research shows that the best way to align interests is to give managers significant stakes in the future value of the company they are running, in the form of restricted stock and option grants. The problem with the current bonus

system is that it rewards short-term targets at the cost of jeopardizing long-run objectives.

Clementi and Cooley (2009) compared the compensation of Wall Street CEOs in the finance, insurance and real-estate sector (FIRE) with pay packages in other sectors. There are two features that stand out. First, executives in the FIRE sector really do make considerably more than their counterparts in other industries. Second, CEOs in the FIRE sector tend to receive a larger part of their compensation as stock (usually restricted stock). In actuality, that structure helps wed executives' interests to those of shareholders.

Problems do, however, emerge when one looks at how companies compute and allocate their bonus pools. Normally, the pools are divided among participating employees according to how much each contributed to the success of the firm. The intent: to reward good past performance and motivate effort in the future.

But all too often the profits that determine the size of the bonus pool are based on trades that produce short-term returns from taking on more risk. For example, in many firms it was enough to book profits on the short-term difference between the yield on AAA-rated mortgage-related securities and the internal cost of funds. This seemed like free money at the time – until these securities turned out to be extremely toxic. Bonuses were based on assets that were not correctly assessed and on profits that were not real. As it turns out, it is not actually possible to tell what 'profits' are, except over a longer time horizon.

Source: adapted from G. C. Clementi and T. F. Cooley, 'Are CEOs paid too much ? How to fix Wall Street's bonus system', *Newsweek*, 5 March (<http://www.newsweek.com/id/187632>).

Adverse selection

Some people smoke cigarettes but others do not. People who smoke reduce their life expectancy. Individuals know whether they themselves smoke, but suppose the insurance company cannot tell the difference and must charge all clients the same premium rate for life assurance.

Suppose the premium is based on mortality rates for the nation as a whole. People who do not smoke know they have an above-average life expectancy and find the premium too expensive. Smokers know their life expectancy is low and realize that the premium is a bargain. Even though the insurance company cannot tell the difference between the two groups, it knows a premium based on the national average will attract only the high-risk people.

One solution is to assume that all clients smoke and charge the correspondingly high premium to all clients. Non-smokers cannot get insurance at what they believe is a reasonable price. They might pay for a medical examination to try to prove

they are low-risk clients who should be charged a lower price. Medical examinations are now compulsory for many insurance contracts.

To check that you understand the difference between moral hazard and adverse selection, say which is which in the following examples. (1) A person with a fatal disease signs up for life insurance. (2) Reassured by the fact that he took out life assurance to protect his dependants, a person who has unexpectedly become depressed decides to commit suicide. (The first was adverse selection, the second moral hazard.)

CONCEPT 12.1

EDUCATION AND SIGNALLING

In Chapter 10 (Case 10.3) we showed that higher education pays off in terms of future earnings. This represents a good reason why students want to invest in higher education. An alternative theory that explains why individuals want to invest in education is the theory of *signalling*. This theory says it could be rational to invest in costly education *even if education adds nothing directly to an individual's marginal product*.

The theory assumes that people are born with different innate ability. Some people are good at most things; other people are less smart and less productive. Not all smart people have blue eyes. The problem for firms is to tell which applicants are the smart ones with high productivity. Looking at their eyes is not enough. There is a problem of asymmetric information.

Suppose higher education contributes nothing to productivity. Signalling theory says that, in going on to higher education, people who know that they are smart send a signal to firms that they are the high-productivity workers of the future. Higher education *screens out* the high-productivity workers.⁵ Firms (the uninformed agents in this case) can pay university graduates more because they send the signal that they are the high-ability workers.

To be effective, the screening process must separate the high-ability workers from the others. Why don't lower-ability workers go to university and fool firms into offering them high wages ? Lower-ability workers may not be confident of passing the necessary exams. If studying adds to productivity, firms should offer higher wages to people who *attend* university, whether or not they pass the final exams. If university screens out the good people, firms will care not about attendance but *academic performance*.

Some firms hire university students before they sit their final exams. Is this evidence refuting signalling theory ? Not necessarily. Screening still works. Since most people know their own ability, firms may take it on trust that people who have stuck it out until their final year at university believe themselves to be at the high end of the ability range.

It seems probable that education (even at the highest level) contributes something to productivity. But there may also be an element of screening.

12.4 Uncertainty and asset returns

There are many ways to carry wealth from the present to the future. People can hold money, government bills or bonds, company shares, housing, gold, and so on. We now compare the rates of return on shares and Treasury bills, two particular ways of holding wealth.

Table 12.3 Average annual real rates of return, 1900–2010 (% per annum)

	2010	1990–2010	1990–2010
Equities	8.9	6.2	5.2
Gilts	4.4	5.9	1.4
Short-term liquid assets	-4.1	2.6	1.0

Source: Barclays Capital, Equity Gilt Study 2011.

Treasury bills are issued usually for a period of three months. The Treasury sells a bill for, say, £99 and simultaneously promises to buy back the bill for £100 in three months' time. People who buy the bill, and later resell it to the government, earn around 1 per cent on their money in three months. By reinvesting the proceeds to buy three more bills in the course of the year, they will earn around 4 per cent a year. Each time an individual buys a bill, the implicit nominal interest rate over the three-month period is known for certain since the government has guaranteed the price at which the bill will be repurchased.

The *real return* is the nominal return minus the inflation rate over the period the bill is held. The real return on Treasury bills is not very risky.

The **capital gain** (loss) is the profit (loss) from selling a share at a price higher (lower) than the purchase price.

Dividends are the regular payments of profit to shareholders.

Company shares offer a return in two different ways. If a share is bought at a low price and later sold at a high price (this is known as a **capital gain**), this contributes to the return earned while holding the share. The *rate of return* is the return as a percentage of the money initially invested. Hence

$$\text{Rate of return} = \frac{\text{dividend} + \text{capital gain}}{\text{initial purchase price}}$$

To compute the real rate of return, we subtract the inflation rate from the nominal rate of return. Table 12.3 compares the average annual real rate of return on company shares (equities) with that on government bonds (gilts) and on short-term assets, such as Treasury bills or interest-bearing liquid loans in the UK during 1900–2010. It shows that, on average, equities yield substantially higher returns.

However, the real rate of return on company shares is much more variable than that on Treasury bills. The latter varied little, but the annual real return on shares was as high as 130 per cent during 1975 and as low as –70 per cent in 1974. There are many years when the real return on shares exceeded 20 per cent or fell below –10 per cent. Shares are much riskier than Treasury bills.⁶ This larger risk is compensated *on average* by a higher return. Since the risk is big – people lost fortunes as shares plummeted after 2007 – it needs a large real return on average to induce people to take this risk.

CASE 12.3

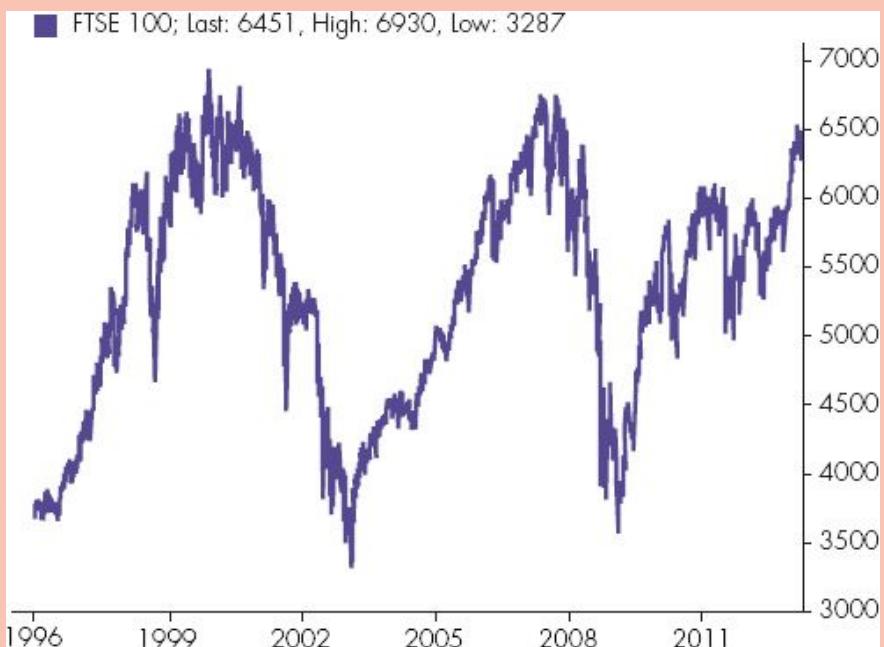
STOCK MARKET VOLATILITY

The chart overleaf shows an index of corporate share prices, the Financial Times Stock Exchange (FTSE) index of 100 top companies from 1996 to 2012. The graph shows that, even for a broad average over 100 individual companies, the index showed many variations over time. Between 1996 and 2000, the index showed an upward trend. In the period 2001–03, the index moved downwards. This downward trend was created by the 11 September terrorist attack in New York and the bursting of the dotcom ‘bubble’. After 2003 the FTSE index started improving – and then came the credit crunch.

The returns from the shares of the companies in the index have varied quite a lot over the last 16 years. In those years inflation in the UK was low and stable, thus implying that *real returns* were very volatile.

Why such wild swings? Largely because the market has to extrapolate current information to make guesses about the entire stream of future earnings of these companies. In volatile sectors, small changes in current information can lead to rapid reassessments about future earnings. Since the share price embodies the future earnings that a company will earn, share prices can change dramatically when uncertainty about a sector is great. When Facebook shares started being traded on the stock market on 18 May 2012, the initial price was \$38 per share. In less than four months the share price plummeted to \$18 per share, a decrease of more than 100 per cent. Why? Investors did not believe that the initial price reflected the real future profitability of Facebook. Towards the end of 2012 news about the good performances of Facebook’s new mobile advertising units and Facebook’s new advertising network

boosted investors' confidence in Facebook's long-term growth prospects. The share price started rising again, reaching \$28.65 in February 2013.



Source: <http://www.digitallook.com>.

Shares are riskier for two reasons. First, nobody is sure what dividend the firm will pay. It depends what profit the firm makes and how confident it is about the future. When firms anticipate tough times, they cut dividends in order to keep a contingency reserve within the firm.

Second, views about the likely capital gains change radically. Stock market investors paid high prices for dotcom companies in the late 1990s, even though profits were still years away. People thought the present value of distant dividends was big. Discounting reduces the value of future dividends, but people were projecting spectacular growth and eventually huge dividends. Growth projections were slashed as reality crept in, and estimated present values changed a lot. Share prices in Amazon and Yahoo! fell by 80 per cent or more during 2000–01. Case 12.3 above gives details over a longer period. Thus, revisions in belief about capital gains are what cause volatile share prices and share returns.

12.5 Portfolio selection

The *portfolio* of a financial investor is the bundle of financial and real assets – bank deposits, Treasury bills, government bonds, shares in industrial companies, gold, works of art – in which wealth is held. How does a risk-averse investor select her portfolio or wealth composition?

Chapter 5 set out the basic model of consumer choice. The budget line summarized the market opportunities – the goods that a given income would buy. Indifference

curves showed individual tastes, and the consumer chose the bundle on the highest possible indifference curve given the budget constraint describing which bundles were affordable.

We use the same approach for the choice of a portfolio. Instead of a choice between different goods, we now focus on the choice between the average or expected return on the portfolio and the risk that the portfolio embodies.

The risk–return choice

Tastes

The risk-averse consumer (or financial investor) prefers a higher average return on the portfolio but dislikes higher risk. To take more risk, she needs to think she will get a higher average return. By ‘risk’ we mean the variability of returns on the *whole portfolio*. From the previous section, we know that a portfolio composed exclusively of industrial shares is much riskier than a portfolio composed only of Treasury bills.

Opportunities

To highlight the problem of portfolio selection, assume there are only two assets in which to invest. Bank deposits are quite a safe asset. Their return is predictable. The other asset is company shares, which are much riskier since their return is more variable.

The investor has a given amount of money to invest. Putting it all in bank deposits, the whole portfolio would earn a small but riskless return. The higher the fraction of the portfolio held in shares, the larger the average return on the whole portfolio but the greater its risk.

Portfolio choice

A very risk-averse investor will put the whole portfolio into the safe asset. To consider buying the risky asset, she must believe the average return on the risky asset is much higher than on the safe asset. Suppose this is the case. How much of the portfolio will she put into the risky asset ? Generally, the fraction of the portfolio held in the risky asset will be higher (1) the higher the average return on the risky asset compared with the safe asset, (2) the less risky is the risky asset, and (3) the less risk-averse is the investor.

Diversification

When there are several risky assets the investor may be able to reduce the risk on the whole portfolio *without* having to accept a lower average return on the portfolio. We illustrate using Table 12.4, whose structure resembles the problem of the actress and the farmer in Table 12.2. There are two risky assets: oil shares and

bank shares. Each has two possible returns: £4 if things go well and £2 if things go badly. Each industry has a 50 per cent chance of good times and a 50 per cent chance of bad times. Finally, we assume that returns in the two industries are independent. Good times in the oil industry tell us nothing about whether the banking industry is having good or bad times.

Table 12.4A diversified portfolio

Oil	Banking	
	Good	Bad
Good	£8	£6
Bad	£6	£4

You have £2 to invest, and oil and bank shares each cost £1. Which portfolio gives the best risk–return combination ? A bank share and an oil share have the same risk and expected return. You are indifferent between buying only oil shares and buying only bank shares. But a superior strategy is to buy one of each and *diversify* the portfolio.

Diversification pools risk across several assets whose individual returns behave differently from one another.

Diversification means not putting all your eggs in one basket. If you put your eggs in one basket, buying, say, two oil shares for your £2, you have a 50 per cent chance of earning £8 and a 50 per cent chance of earning £4. It depends on whether the oil industry has good times or bad times. The average return is £6, but the actual return will either be £4 or £8.

Table 12.4 shows a diversified portfolio with one bank share and one oil share. If both industries do well, you will make £8, but this is only a 25 per cent chance. There is a 50 per cent chance of oil doing well; since returns in the two industries are independent, on only half of those occasions will banking also be doing well. Similarly, there is a 25 per cent chance of both industries doing badly at the same time. There is also a 25 per cent chance that one industry does well while the other one does badly. Each of the four portfolio returns shown in Table 12.4 is a 25 per cent chance.

The average return on the portfolio is still £6, as if you put your £2 in one basket, but the variability of returns is smaller. Instead of a 50/50 chance of £4 or £8, you now have only a 25 per cent chance of each extreme, and a 50 per cent chance of earning the average return of £6.

Diversification reduces the risk by pooling it without altering the average rate of return. It offers you a better deal. As in our earlier discussion of risk-pooling by insurance companies, the greater the number of risky assets with independent

returns across which the portfolio pools the risk, the lower will be the total risk of the portfolio.

Figure 12.1 shows the typical relationship between total portfolio risk and the number of independent assets in the portfolio. Portfolio risk declines as the number of independent risky assets is increased. However, most of the risk reduction through diversification comes very quickly. Even a few assets cut the total risk a lot. Your car has one spare tyre, not five.

Because it is more expensive to buy in small quantities, small investors typically hold a dozen different shares rather than a hundred. They get most of the benefit of diversifying without needing lots of small packages of shares. People who are more risk-averse and want a large number of shares can buy shares in a mutual fund or unit trust, a professional fund that buys large quantities of shares in many firms, and then retails stakes in the fund to small investors.

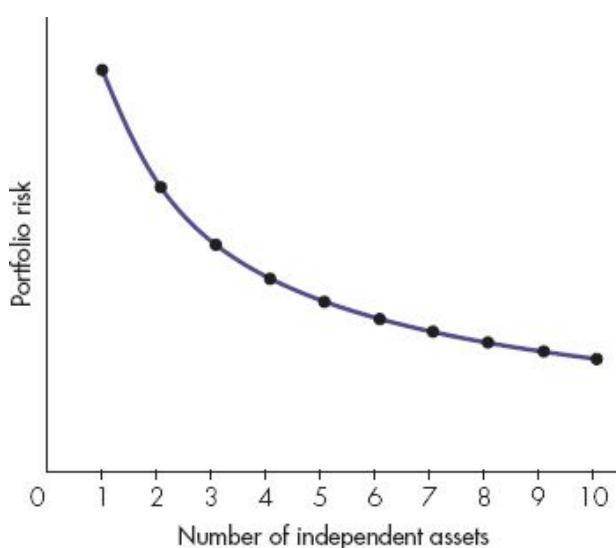


Figure 12.1 Diversifying risk

The riskiness of a portfolio, the variability of its total return, can be minimized through diversification, but the gains to further diversification diminish quite rapidly.

Diversification when asset returns are correlated

Risk pooling works because asset returns are independent of each other. When asset returns move together, we say that they are *correlated*. When returns on two assets tend to move in the same direction, we say they are *positively correlated*. For example, a boom in the whole economy will tend to be good for bank shares and shares in TV companies. If returns tend to move in opposite directions, we say they are *negatively correlated*. For example, if people buy gold shares during financial crises, gold shares will tend to rise when other shares are falling, and vice versa.

Positive and negative correlations affect the way in which diversification changes risk. Suppose bank shares and oil shares always rise or fall together. Buying one of each is like putting all your money in either share. Diversification achieves nothing. When returns are perfectly positively correlated, risk pooling does not work, just as it fails for ‘acts of God’ in the insurance industry.

Conversely, diversification is a spectacular success when returns are negatively correlated. Suppose bank shares do well only when gold shares do badly, and vice versa. Buying one of each, you earn either £4 from oil and £2 from gold or £2 from oil and £4 from gold. With the diversified portfolio, you earn £6 for certain. You have diversified away all the risk, even though each share is individually risky.

In practice, returns on different shares are never perfectly correlated. Some *tend* to vary together and some *tend* to vary in opposite directions, but over any particular period actual returns on two shares may not exhibit their usual correlation. Thus it is impossible to completely diversify away all portfolio risk. But smart fund managers are always on the lookout for an asset that tends to have a negative correlation with the assets in the existing portfolio. On average, extending the portfolio to include that asset will improve the risk–return characteristics of the portfolio.

Beta

Table 12.5 gives some examples of **beta**. The first row shows returns on the market as a whole in booms, normal times and slumps. A share with beta 5 1 moves the same way as the whole market. A high beta share does even better when the market is up, even worse when the market is down. A low beta share moves in the same general direction as the market but more sluggishly than the market. Negative beta shares move against the market.

Beta measures how much an asset’s return moves with the return on the whole stock market.

Table 12.5 Share returns and beta

Asset	Return (%)		
	Boom	Normal	Slump
Whole market	14	6	22
High beta	20	10	28
Beta = 1	14	6	22
Low beta	5	4	3
Negative beta	2	3	5

Most shares move pretty much with the market and have a beta close to unity. There are not too many negative beta shares, but some gold shares have betas close to zero. Most people should have some gold shares in their portfolios.

Bankers and stockbrokers calculate betas from the past behaviour of individual shares and the whole stock market. Ideally, they are looking for negative beta shares that greatly reduce the risk of a portfolio whose other components vary with the market as a whole. Even low beta shares are partly independent of the rest of the market and allow some risk to be pooled. High beta shares are undesirable. Including them in the portfolio adds to its total risk.

A share with a low (or even negative) beta will be in high demand. Risk-averse purchasers are anxious to buy low beta shares whose inclusion in their portfolios reduces the total portfolio risk. High demand bids up the share price and reduces the average return: since it costs more to buy the shares, people get fewer per pound invested. However, investors are happy to trade off a lower return for the fact that low beta shares reduce the total risk of their portfolios.

In stock market equilibrium, low beta shares have high prices and low rates of return on average. Conversely, high beta shares add to investors' portfolio risk and are purchased only because they have low prices and on average offer high rates of return that compensate for their undesirable risk characteristics. Figure 12.2 shows the results of a pioneering study by Professors Black, Jensen and Scholes⁷ using stock market data from 1931 to 1965. Average returns on individual shares rise steadily with the shares' beta, as the theory predicts. Table 12.6 shows recent estimates of beta for selected sectors in the US.

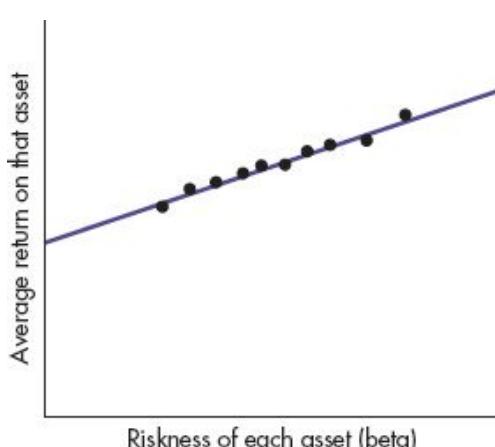


Figure 12.2 Global issuance of bonds backed by mortgages

Each share's risk is measured by its beta, which shows how that share's returns move with returns in the market as a whole. The higher the beta, the more the inclusion of the share in a portfolio will increase the total portfolio risk. The data show that riskier shares with higher betas must offer a higher return on average to compensate for this disadvantage.

Table 12.6 Beta for selected sectors in the US (January 2013)

Retailing	1.28	Electricity	0.52
Computer software	0.98	IT services	1.05
Banks	0.77	Automotive	1.73
Chemicals	1.36	Household products	0.98
Beverages	0.95	Internet	1.17
E-commerce	1.05	Steel	1.65
Tobacco	0.86	Aerospace/Defense	0.98

Source:

http://people.stern.nyu.edu/adamodar/New_Home_Page/datafile/Betas.html.

To sum up, individual share prices depend both on expected or average returns and on risk characteristics. The risk characteristics of a firm's shares determine the expected return its shares must offer to compete with other shares. For a given required return, higher anticipated income (dividends or capital gains) means a higher current share price.

The riskiness of a firm's shares refers not to variability of the share's return in isolation from the rest of the market. This is why beta matters. Adding a risky asset to the portfolio reduces the risk of the portfolio provided the share's beta is less than 1. Low beta shares can be individually risky; nevertheless, taken with other shares they reduce portfolio risk and are therefore desirable. Low beta shares have an above-average price and a below-average rate of return to offset this advantage; high beta shares must offer an above-average expected return to be competitive.

ACTIVITY 12.1

BETA IN ACTION

The table below shows three possible outcomes – boom, slump and normal times – and three possible assets – the FTSE index, an asset with negative beta that moves against the general stock market trend, and a high beta asset that moves in the same direction as the stock market but even more so.

Outcome	Asset price			Portfolio value		
	(a) FTSE index	(b) Low beta asset	(c) High beta asset	A $\frac{1}{2}$ of (a) + $\frac{1}{2}$ of (b)	B $\frac{1}{3}$ of (a) + $\frac{2}{3}$ of (b)	C $\frac{1}{2}$ of (a) + $\frac{1}{2}$ of (c)
Boom	120	90	150			
Normal	100	100	100			
Slump	80	110	50			

Questions

- Complete the table by showing for each outcome (boom, normal, slump) the value of portfolios A, B and C, respectively (for example, in portfolio A half of your money is invested in the FTSE index and half in the low beta asset).

To check your answers to this question, go to page 678.

Diversification in other situations

Risk is all around us, and diversification happens all the time. Countries diversify their sources of raw materials; otherwise, if anything disrupts the sole supplier's ability or willingness to sell, the country may face a disaster. Similarly, a farmer is reluctant to rely on a single crop. It may be better for a navy to have two small aircraft carriers than one large one. If the only aircraft carrier sinks, there is no air cover.

12.6 Efficient asset markets

An **efficient asset market** already incorporates existing information properly in asset prices.

There are two basic images of the stock market. One is that of a casino, without any rational basis for speculation; it is all a matter of luck. The other view – the theory of **efficient markets** – is that the stock market is a sensitive processor of information, quickly responding to new information to adjust share prices correctly.

The second view recognizes that share prices fluctuate a lot but argues that these fluctuations are the appropriate response to new information as it becomes available.

Companies with high average returns and low betas should be valued both by society and by the stock market. The higher the share price, the more money a company raises from a new share issue, and the more likely is the company to invest in plant and machinery financed by this new share issue. High share prices are guiding the right firms to invest. Companies with low average returns and high betas are valued neither by financial investors nor by society. Low share prices make it harder for them to finance new plant and equipment, and they will tend to contract.

It matters which of the two views of the stock market is correct. If share prices correctly reflect prospective dividends and risk characteristics – the efficient market view – a free market in industrial shares is guiding society's scarce resources towards the right firms. But if share prices are purely pot luck, as in a casino, the wrong firms may expand just because their share prices are high.

Testing for efficiency

Suppose everybody has all the information available today about the likely risks and returns on different shares. Equilibrium share prices should equate the likely return on all shares with the same risk characteristics. Otherwise there would be an obvious opportunity to switch from the low return shares to higher return shares with equivalent risk characteristics. If the market has got it right, it does not matter which share you buy in any risk class. They are all expected to yield the same return. The efficient market view says there is no way of beating the market to earn an above-average return on a share of a given risk class.

If the market neglects some available information, you could use this information to beat the market. If the market failed to spot that hot weather increases ice cream sales, it would never mark up share prices in ice cream companies when good weather occurred. By buying ice cream shares when the sun shone you would make money and beat the market. The market would be surprised by high dividends from ice cream companies. But you bought them, having figured all this out by using extra information. You knew ice cream shares would pay a higher rate of return than the market thought. You spotted an inefficiency in the market.

In contrast, the efficient market view says all the relevant available information is immediately incorporated in the share price. Given the long-range weather forecast, the market makes the best guess about profits and dividends in the ice cream industry and sets the current price to give the required rate of return for shares with the same risk characteristics as ice cream shares. If the weather forecast is correct, the return will be as predicted. If there is an unexpected hot spell, the market will immediately mark up ice cream shares to reflect the new information that ice cream profits will be higher than previously expected. How high are ice cream shares marked up ? To the price that reduces the expected rate of return back to the average for that risk class.

The crucial implication of the efficient market theory is that asset prices correctly reflect all existing information. It is unforeseen new information that changes share prices as the market quickly incorporates this unanticipated development to restore expected returns to the required level. Existing information cannot systematically be used to get above-average returns for that risk class of asset.

The theory of efficient markets has been tested extensively to see whether there is any *currently available* information that would allow an investor systematically to earn an above-average return for that risk class. The vast majority of all empirical studies conclude that there is no readily available information that the market neglects. Rules of the form ‘buy shares when the price has risen two days in a row’ do not work. Nor do rules that use existing information about how the economy or the industry is doing. Smart investors have taken this information on board as it became available. It is already in the price.

The empirical literature usually concludes that you may as well stick a pin in the financial pages of a newspaper as employ an expensive financial adviser. Paradoxically, it is because the market has *already* used all the relevant economic information correctly that there are no bargains around. The theory of efficient

markets does not say share prices and returns are unaffected by economics; it says that, because the economics has been correctly used to set the price, there are no easy pickings left.

Financial newspapers and stock market institutions run competitions for the investor of the year. If the theory of efficient markets is right, why do some portfolios do better than others? Why, indeed, are financial portfolio advisers in business at all? The world is uncertain, and there will always be surprises that could not have been forecast. As this new information is incorporated in share prices, some lucky investors will find they happen to have already invested in shares whose price has unexpectedly risen. Others are unlucky, holding shares whose price unexpectedly falls.

Thus one interpretation of why some investors do better than others is pure chance. This story could even explain why some investors have above-average returns for several years in a row. Even with a fair coin there is roughly one chance in a thousand of tossing ten consecutive heads. Even if there is no systematic way to beat the market, there are thousands of investors, and someone is going to have a lucky streak for ten years.

But there is also a more subtle interpretation. When a piece of new information first becomes available, someone has to decide *how* share prices should be adjusted. The price does not change by magic. And there is an incentive to be quick off the mark. The first person to get the information, or to calculate correctly where the market will soon be setting the price, may be able to buy a share just before everyone else catches on and the share's price rises.

The non-specialist investor cannot use *past* information to make above-average profits. But specialist investors, by reacting very quickly, can make capital gains or avoid capital losses within the first few hours of new information becoming available. It is their actions that help to change the price, and the profits that they make from fast dealing are what pay for City salaries. It is the economic return on their time and effort in gathering and processing information.

Speculative bubbles

Consider the market for gold. Unlike shares or bonds, gold pays no dividend or interest payment. Its return accrues entirely through the capital gain. Today's prices depend on the anticipated capital gain, which in turn depends on expectations of tomorrow's price. But tomorrow's price will depend on the capital gain then expected, which will depend on expectations of the price the day after; and so on.

In such markets there is no way for the *fundamentals*, the economic calculations about future dividends or interest payments, to influence the price. It all depends on what people today think people tomorrow will expect people the next day to expect. Such a market is vulnerable to *speculative bubbles*. If everyone believes the price will rise tomorrow, it makes sense to purchase the asset today. So long as

people expect the price to keep rising, it makes sense to keep buying even though the price may already have risen a lot.

A famous example of a speculative bubble is the South Sea Bubble of 1720. A company was set up to sell British goods to people in the South Seas and to bring home the wonderful and exotic goods produced there. The shares were issued long before any attempt was made to actually trade these goods. It sounded a great idea and people bought the shares. The price rose quickly, and soon people were buying not in anticipation of eventual dividends but purely to resell the shares at a profit once the price had gone even higher. The price rose even faster, until one day it became apparent that the company's proposal was a fiasco with no chance of success and the bubble burst. Sir Isaac Newton lost £2000 (over half a million pounds at today's prices).

The great English economist John Maynard Keynes argued that the stock market is like a casino, dominated by short-term speculators who buy not in anticipation of future dividends but purely to resell at a quick profit. Since next period's share price depends on what people then think the following period's share price will be, Keynes compared the stock market to a beauty competition in a newspaper, where the winner is the reader who guesses the beauty receiving most votes from all readers. Share prices reflect what average opinion expects average opinion to be.

Undoubtedly, financial markets sometimes exhibit temporary bubbles. The most recent example was the housing market bubble in the US market that lead to the credit crunch of 2007–08. Nevertheless, bubbles *are* usually temporary. Eventually it is obvious that the share price cannot be justified by fundamentals. Bubbles are less likely for assets whose income is mainly from dividends or interest rather than capital gains.

CASE 12.4

EXCESSIVE RISK AND THE 2008 FINANCIAL CRISIS: A BEHAVIOURAL VIEW

At the heart of the recent financial crisis was the excessive risk taken by traders in assets related to sub-prime mortgages. Why did traders take so much risk?

Asymmetric information has certainly played a role. Traders may not have been fully aware of the quality of the assets they were investing in (adverse selection). Moreover, large investment banks may have believed that, even if their investments were going to crash, governments would step in to help them, thus avoiding the collapse of the entire financial system (moral hazard). Those two facts may explain why excessive risk was taken. However, this is probably just a part of the overall story.

Behavioural economics and finance may offer a different view on why traders took too much risk. This branch of economics investigates how psychology may affect individuals' decisions. An example of how psychology may affect excessive risk behaviour is *cognitive dissonance*. Cognitive dissonance is the discomfort we feel when we take an action that conflicts with our typically positive self-image. Research has shown that smokers often experience cognitive dissonance. A smoker knows that he is doing something harmful to his health. However, he may reduce the discomfort he feels by convincing himself that smoking is not, after all, as risky as some say. He may, for example, remind himself of other smokers who, despite smoking, seem to be doing just fine. The smoker manipulates his beliefs to reduce his discomfort.

A similar story can be applied to traders in sub-prime mortgages. If a trader begins to sense that his investment in sub-prime assets may create serious risks to his institution and to the broader financial system, this will threaten his positive self-image of a person whose work is valuable. This may create uncomfortable dissonance. To reduce this dissonance, he could resign, but that action is financially costly for him. Instead, he can manipulate his beliefs, telling himself that his investments are not that risky. He might stop himself from inspecting too closely the quality of the sub-prime loans and stumble on some disturbing information. By doing so, he will continue to invest excessively and take too much risk.

Source: N. Barberis, 'Psychology and the financial crisis 2007–2008', working paper (<http://ssrn.com/abstract=1742463>).

12.7 More on risk

A **forward market** deals in contracts made today for delivery of goods at a specified future date at a price agreed today.

A **spot market** deals in contracts for immediate delivery and payment.

Risk is central to economic life. Every topic in this book could be extended to include risk. Individual applications differ, but two features recur: individuals try to find ways to reduce risk, and those who take over the bearing of risk have to be compensated for so doing.

Hedging and forward markets

There are **forward markets** for many commodities and assets, including corn, coffee, sugar, copper, gold and foreign currencies.

Suppose the current price of copper is £800 a tonne and people expect the price to rise to £880 a tonne after 12 months. Some people will hold copper in their portfolios. The expected capital gain is 10 per cent of the purchase price, and it may be interesting to diversify a portfolio by including copper. However, that is not our concern at present.

You own a copper mine and will have 1 tonne of copper to sell in 12 months' time. The **spot** price of copper is the price for immediate delivery. Today's spot price is £800 and people expect the spot price to be £880 at this time next year. One option is for you simply to sell your copper at the spot price at this time next year. You expect that to be £880 but you cannot be sure today what the price next year will actually be. It is risky.

A **Hedging** is the use of forward markets to shift risk on to somebody else.

Alternatively, you can **hedge** against this risk in the forward market for copper. Suppose today you can sell 1 tonne of copper for delivery in 12 months' time at a price of £860 agreed today. You have hedged against the risky future spot price. You know for certain what you will receive when your copper is available for delivery. But you have sold your copper for only £860, even though you expect copper then to sell for £880 on the spot market. You regard this as an insurance premium to remove the risk associated with the future spot price.

A **speculator** temporarily holds an asset in the hope of making a capital gain.

To whom do you sell your copper in the forward market? You sell it to a trader, whom we can call a **speculator**. The speculator has no interest in 1 tonne of copper *per se*. But the speculator, having promised you £860 for copper to be delivered in one year's time, currently expects to resell that copper immediately it is delivered. The speculator expects to get £880 for that copper in the spot market next year. He expects to make £20 as compensation for bearing your risk. If spot copper prices turn out to be less than £860 next year, the speculator will lose money. £20 is the risk premium necessary to attract enough speculators into the forward market to take up the risky positions that hedgers wish to avoid.

Someone buying spot copper today at £800 for possible resale next year at £880 must compare the expected capital gain of 10 per cent with returns and interest rates on offer in other assets. Copper must cover the opportunity cost of the returns that could have been earned by using this money elsewhere. The speculator in the forward market need not make this comparison. No money is currently tied up in the forward contract. Although the price has been agreed today at £860, the money is handed over only next year when the copper is delivered. Provided the speculator

then resells in next year's spot market, no money is actually tied up. All the speculator has to think about is the likely spot price in 12 months' time and how much it could vary either side of this estimate. The riskier the future spot price, the larger premium the speculator will need and the more the current forward price will lie below the expected future spot price. All the information is summarized in Table 12.7.

Table 12.7 Summary of the spot and forward markets for copper

Today's spot price	Price of copper today for delivery and payment today	£800
Future spot price	Spot market price of copper in a year's time	£900
Expected future spot price	The best guess today about spot price in a year's time	£880
Forward price	Price today in forward market at which copper is being traded for delivery and payment in a year's time	£860
Risk premium	Expected future spot price minus the current forward price. The sum a hedger expects on average to lose by making a forward contract rather than by taking a chance on the future spot price. Hence, what the hedger expects to pay, and the speculator expects to make, by transferring the risk from hedger to speculator	£880 - £860 = £20

This speculator had an open position, having taken forward delivery of copper without yet having a purchaser to whom to resell. However, other firms use copper as an input to production, and may wish to *buy* copper for delivery in 12 months' time at a price agreed today. They too wish to hedge against the risky future spot price. A speculator who can make two forward contracts, one to take delivery of copper from the copper miner, the other to sell copper to a copper user, does not have an open position. The speculator's book is balanced, without any residual risk. The risky future spot price is irrelevant.

In forward markets with roughly equal numbers of people wishing to hedge by buying and by selling, speculators' books roughly balance and residual risk is small. Speculators need only a little compensation to cover this residual risk and the administration costs. The current price of forward copper is close to the expected future spot price.

However, speculation is a risky business if buyers and sellers cannot be matched up in the forward market. In practice, the spot prices that subsequently transpire can vary by a large amount on either side of the estimate implicitly contained in the current forward price.

Why do forward markets exist for copper and silver but not for BMWs? The answer again is moral hazard and adverse selection. Suppose today you contract for delivery of a new car model in 12 months' time. You thought you were buying a luxury car, but the company brings out a low-quality car and says, 'This is our new model'. By making all these forward contracts, the car maker affects its own quality incentives.

Forward markets do not exist for most goods because it is impossible to write legally binding and cheaply enforceable contracts that adequately specify the characteristics of the commodity being traded. Where forward markets exist they are for very standardized commodities – 18-carat gold, Japanese yen – that are easily defined. Forward markets are an important way in which individuals can reduce the risks they face, but there are only a limited number of risks that can be hedged in this way.

Compensating differentials in the return to labour

Since people are risk-averse, we expect those with risky jobs to earn more on average than people whose jobs are safe. Broadly speaking, this is confirmed by the facts. Divers who inspect North Sea oil pipelines earn high hourly rates because the death rate in this activity is high. University academics earn relatively low wages in the UK because many of them have secure jobs, unlike industrial managers who face the sack if their company has a bad spell.

Profits are often seen as a reward to entrepreneurs, individuals who set up and run firms, for taking big risks. The average person who starts a business initially works long hours for small rewards. In the early stages there is the continual threat of failure, and most small firms never get off the ground. The possibility of becoming a millionaire, like Richard Branson of Virgin or Bill Gates of Microsoft, is the carrot needed to persuade people to embark on this risky activity.

Summary

- **Risk** pervades economic life. Some people gamble for fun; some addicts gamble in spite of themselves. Most people are **risk-averse**. They volunteer to take risks only if offered favourable odds that on average yield a profit. Conversely, most people **insure**, despite less than fair odds, to reduce the risks they otherwise face.
- **Risk-aversion** reflects the **diminishing marginal utility of wealth**. A fair gamble in monetary terms yields less extra utility when it succeeds than it sacrifices when it fails. Hence people refuse fair gambles, except for very small stakes. The prevalence of risk aversion means that people look for ways to reduce risk, and must pay others to take over their risk bearing.
- Insurance **pools** risks that are substantially independent to reduce the aggregate risk, and **spreads** any residual risk across many people so that each has a small stake in the risk that cannot be pooled away.
- Insurance markets are inhibited by **adverse selection** and **moral hazard**. The former means that high-risk clients are more likely to take out insurance; the

latter means that the act of insuring increases the likelihood that the undesired outcome will occur.

- Company shares have a higher average return but a much more variable return than that on Treasury bills or bank deposits.
- Portfolio choices depend on the investor's tastes – the trade-offs between risk and average return that yield equal utility – and on the opportunities that the market provides – the risk and return combinations on existing assets.
- When risks on different asset returns are independent, the risk of the whole portfolio can be reduced by **diversification** across assets.
- The risk that an asset contributes to a portfolio is not measured by the variability of that asset's own return but by the correlation of its return with the return on other assets. An asset that is negatively correlated with other assets will actually reduce the risk of the whole portfolio even though its own return is risky. Conversely, assets with a strong positive correlation with the rest of the portfolio increase the overall risk. The value of **beta** for an asset measures its correlation with other assets.
- In equilibrium risky assets earn higher rates of return on average to compensate portfolio holders for bearing this extra risk. High beta assets have high returns. If an asset is offering too high an expected return for its risk class, people will buy the asset, bidding up its price until the expected return is forced back to its equilibrium level.
- In an **efficient market** assets are priced to reflect the latest available information about their risk and return. There are no easy systematic investment opportunities to beat the market unless you systematically get or use new information faster than other people. Evidence from share prices is compatible with stock market efficiency, but speculative bubbles sometimes occur.
- **Forward markets** set a price today for future delivery of and payment for goods. They allow people to **hedge** against risky **spot** prices in the future by making a contract today. **Speculators** take over this risk and require a premium unless they can match buyers and sellers.

Review questions



EASY

- 1 A fair coin is to be tossed. If it comes down heads, the player wins £1. If it comes down tails, the player loses £1. Person A doesn't mind whether or not she takes the bet. Person B will pay £0.02 to play the game. Person C demands £0.05 before being willing to play. Characterize the three people's attitudes to risk. Which person is most likely to take out insurance against car theft ?
- 2 In which of the following are the risks being pooled: (a) life insurance, (b) insurance against the Thames flooding, (c) insurance for a pop star's voice ?
- 3 You see an advert for life insurance for anyone over 45 years old. No medical examination is required. Do you expect the premium rates to be high, low or average ? Why ?
- 4 **True or False** Your lecturer says that the economics exam she is setting is so easy that no students will fail. Given that information, no students prepare for the exam and they all fail. This is an example of an adverse selection problem.
- 5 **Common fallacies** Why are the following statements wrong ? (a) Economists cannot predict changes in the stock market. This proves that economics is useless in thinking about share prices. (b) It is silly to take out insurance. If the insurance company is making money, its clients are losing money. (c) Prudent investors should not buy shares whose returns are volatile.

MEDIUM

- 6 Suppose that Philip has a total utility of income given by $U(I) = I$, where I denotes income. In a graph with utility on the vertical axis and income on the horizontal axis, plot Philip's total utility of income. What is Philip's attitude towards risk ? Explain.
- 7 Suppose that Philip, whose total utility of income is given by $U(I) = I$, can have two income levels this year: if he keeps his job, he earns £30 000; if he loses his job, he must live on a subsistence payment from the government of £10 000. Philip thinks he has a 50–50 chance of losing his job when he talks to his boss tomorrow. He is thus considering purchasing employment insurance today. The insurance company will pay him £20 000 if he loses his job; otherwise, it will pay nothing. The insurance payment does not affect the subsistence payment from the government. The insurance costs £10 000, which he must pay now. Using the graph from Question 6, show the possible income levels Philip faces. Should Philip buy the insurance ? Explain.
- 8 We know from many situations that people will pay to avoid risk. Name three risky products that you could choose to buy. In each case, explain the motive.
- 9 What kind of information asymmetry do stock markets face which requires that these markets be regulated ?
- 10 You set up a firm to advise the unemployed on the best way to use their time to earn money. Your firm issues shares on the stock market. In equilibrium, will your shares be expected to earn a higher or lower return than the stock market average ? Why ?
- 11 Suppose the stock exchange is expected to yield a return of 5 per cent next year, but this is risky and could be several percentage points either side of the central forecast. You are also aware that it is possible to hold gold as an asset and that

gold is known to have a small negative beta. People buy gold in a panic, so the gold price rises when the stock market is doing badly. Today, the price of gold is £100. (a) If people are risk-neutral, what is the best estimate of next year's gold price ? (b) If people are risk-averse, what do you think is the best estimate of next year's gold price ?

HARD

- |2 The total utility of the income of an individual is reported in the following table:

Income	Utility
1	1.00
2	1.26
3	1.44
4	1.59
5	1.71
6	1.82
7	1.91
8	2.00
9	2.08
10	2.15

In a graph with utility on the vertical axis and income on the horizontal axis, plot the total utility of income of the individual. Suppose our individual has an initial income of £4. He can use this money in a fair gamble that gives him an income of £6, with probability 0.5, and with probability 0.5, an income of £2. Should the individual put his income into this fair gamble ? Use the graph of the total utility of income to illustrate your answer.

- |3 **Essay question** You run a pension fund and know that in 20 years' time you need to make a lot of payments to people who will then have retired. Should you: (a) invest in bonds that mature in 20 years' time so you know exactly how much you will then have, (b) invest in equities because historically their average return has been greater than bonds in the long run, or (c) begin mainly in equities but switch gradually into bonds as the 20-year period elapses ?
- |4 Suppose you have 1 tonne of coffee to be sold in a year's time. The spot price of coffee today is £1000. The spot price of coffee is expected to be £1300 next year. How would you protect yourself from risk and make the most out of this trade ?

1 This is the 'law of large numbers'. Proof of this law can be found in most statistics textbooks.

2 Uncertainty and risk are not exactly synonymous. Uncertainty refers to situations where the probability of certain occurrence is not known. When uncertainty can be quantified, we call it risk. However, many people

do not distinguish between the two and we take the same approach.

- 3 In Chapter 5 we discussed the case in which utility depends only on consumption. Here, you can consider the case where utility depends on income, since income is ultimately used for consumption. Notice that in Chapter 5 utility was ordinal. Here, instead, utility of income is cardinal (see the Appendix to Chapter 5).
- 4 Asymmetric information will also be discussed in Chapter 13, where we show how it relates to the problem of market failure.
- 5 Screening is the process of learning inside information by observing differences in behaviour.
- 6 Large positive or negative returns on shares were probably not forecast by the market. If people had foreseen a real return of 30 per cent they would have bought shares earlier, bidding up share prices earlier. If large capital losses had been foreseen, share prices would already have been lower, as people tried to dump these shares before they fell.
- 7 F. Black, M. C. Jensen and M. Scholes, ‘The capital asset pricing model: Some empirical tests’, in M. C. Jensen (ed.), *Studies in the theory of capital markets* (Praeger, 1972), pp. 79–121.

PART THREE

Welfare economics

Normative or welfare economics is concerned with making value judgements and using these to recommend which policies are desirable. Much of economics is about reconciling the goals of efficiency and fairness. Part Three discusses reasons for market failures that give rise to inefficiencies and then investigates how government might intervene to improve the market. Such intervention may itself be subject to failure: well-meaning intervention can sometimes make things worse. As globalization begins to undermine the economic sovereignty of nation states, it is also necessary to think about when national policies will suffice and when cross-border co-operation is beneficial.

Chapter 13 introduces welfare economics, defines efficiency and equity (fairness), and examines reasons for market failure. Chapter 14 focuses on direct government intervention through taxes and public spending.

Contents

- 13 Welfare economics**
- 14 Government spending and revenue**

CHAPTER 13

Welfare economics

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 understand what we mean by welfare economics
- 2 describe horizontal and vertical equity
- 3 understand the concept of Pareto efficiency
- 4 recognize how the ‘invisible hand’ may achieve efficiency
- 5 define the concept of market failure
- 6 recognize why partial removal of distortions may be harmful
- 7 identify the problem of externalities and possible solutions
- 8 understand how monopoly power causes market failure
- 9 analyse distortions from pollution and congestion
- 10 understand why missing markets create distortions
- 11 analyse the economics of climate change

In this chapter we analyse the concepts of efficiency and equity (fairness), and examine reasons for market failure.

Welfare economics deals with normative issues. It does not describe how the economy works but assesses how well it works.

Horizontal equity is the identical treatment of identical people.

Vertical equity is the different treatment of different people in order to reduce the consequences of these innate differences.

Chapter 1 noted that markets are not the only way society can resolve what, how and for whom to produce. Communist economies relied heavily on central direction or command. Are markets a good way to allocate scarce resources? What is a ‘good’ way? Is it fair that some people earn much more than others in a market economy? These are not positive issues about how the economy works but

normative issues about how well it works. They are normative because the assessment depends on the value judgements adopted by the assessor.

Left – and right-wing parties disagree about how well a market economy works. The right believes the market fosters choice, incentives and efficiency. The left emphasizes the market's failings and the need for government intervention. What lies behind the disagreement? Two themes recur in the analysis of **welfare economics**. The first is *allocative efficiency*. Is the economy getting the most out of its scarce resources or are they being squandered? The second is *equity*. How fair is the *distribution* of goods and services among different members of society?

13.1 Equity and efficiency

Whether or not either concepts of equity – horizontal or vertical – are desirable is a pure value judgement. **Horizontal equity** rules out discrimination between people whose economic characteristics and performance are identical. **Vertical equity** is the Robin Hood principle of taking from the rich to give to the poor.

Many people agree that horizontal equity is a good thing. In contrast, the extent to which resources should be redistributed from the ‘haves’ to the ‘have-nots’ to increase vertical equity is an issue on which people disagree.

A **resource allocation** is a complete description of who does what and who gets what.

Efficient resource allocation

Suppose that **allocations** are made by a dictator. Feasible allocations depend on the technology and resources available to the economy. The ultimate worth of any allocation depends on consumer tastes – how people value what they are given.

Figure 13.1 shows an economy with only two people, David and Susie. The initial allocation at *A* gives David a quantity of goods Q_D and Susie a quantity Q_S . Are society's resources being wasted? By reorganizing things, suppose society can produce at *B*, to the north-east of *A*. If David and Susie assess utility by the quantity of goods they get themselves, and if they would each rather have more goods than less, *B* is a better allocation than *A*. Both David and Susie get more. It is inefficient to produce at *A* if production at *B* is possible. Similarly, a move from *A* to *C* makes both David and Susie worse off. If it is possible to be at *A*, it is inefficient to be at *C*.

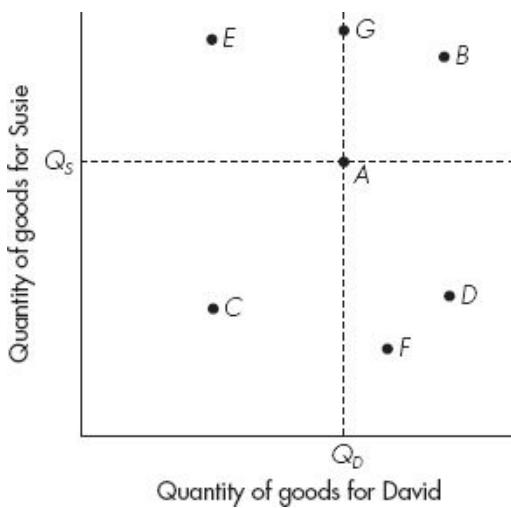


Figure 13.1 Allocating goods to two people

Provided people assess their own utility by the quantity of goods that they themselves receive, B is a better allocation than A, which in turn is a better allocation than C. But a comparison of A, with points such as D, E or F, requires us to adopt a value judgement about the relative importance to us of David's and Susie's utility.

What about a move from *A* to *E* or *F*? One person gains; the other person loses. Whether this change is desirable depends on how we value David's utility relative to Susie's. If we think David's utility is very important we might prefer *F* to *A*, even though Susie's utility is reduced.

Since different people will make different value judgements, there is no unambiguous answer to the question of whether a move from *A* to *D*, *E* or *F* is desirable. It depends on who makes the assessment.

For a given set of consumer tastes, resources and technology, an allocation is **Pareto-efficient** if there is no other feasible allocation that makes some people better off and nobody worse off.

To try to separate the discussion of equity from the discussion of efficiency, modern welfare economics uses the idea of **Pareto efficiency**, named after the economist Vilfredo Pareto.

In Figure 13.1 a move from *A* to *B* or *A* to *G* is a *Pareto gain*. Susie is better off; David is no worse off. If *B* or *G* is feasible, *A* is *Pareto-inefficient*. A free lunch is available.

A move from *A* to *D* makes David better off but Susie worse off. The Pareto criterion has nothing to say about this change. To evaluate it, we need a judgement about the relative values of David's and Susie's utility. The Pareto principle is of limited use in comparing allocations on efficiency grounds. It only allows us to

evaluate moves to the north-east or the south-west in Figure 13.1. Yet it is the most we can say about efficiency without making value judgements about equity.

Figure 13.2 takes the argument a stage further. By reorganizing production, we can make the economy produce anywhere inside or on the frontier AB . From inside the frontier, a Pareto gain can be achieved by moving to the north-east on to the frontier. Any point inside the frontier is Pareto-inefficient. One person can be made better off without making the other worse off. But *all* points on the frontier are Pareto-efficient. One person can get more only by giving the other person less. Since no Pareto gain is possible, every point on the frontier is Pareto-efficient.

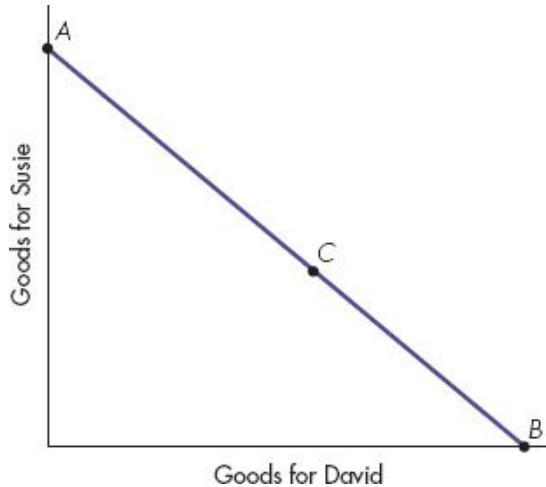


Figure 13.2 The efficient frontier

The frontier AB shows the maximum quantity of goods which the economy can produce for one person given the quantity of goods being produced for the other person. All points on the frontier are Pareto-efficient. David can only be made better off by making Susie worse off, and vice versa. The distribution of goods between David and Susie is much more equal at point C than at points A or B.

Thus society should never choose an inefficient allocation inside the frontier. Which of the efficient points on the frontier is most desirable will depend on the value judgement of the relative values of David's and Susie's utility, a judgement about equity.

13.2 Perfect competition and Pareto efficiency

Will a free market economy find a Pareto-efficient allocation, or must it be guided there by government intervention?

Competitive equilibrium in free markets

Suppose there are many producers and many consumers, but only two goods: meals and films. Each market is a free, unregulated market and is perfectly competitive. In equilibrium, suppose the price of meals is £5 and the price of films is £10. Labour is the only factor of production and workers can move freely between industries. We now work through seven steps:

1. The last film yields consumers £10 worth of extra utility. If it yielded less (more) extra utility than its £10 purchase price, the last consumer would buy fewer (more) films. Similarly, the last meal must yield consumers £5 worth of extra utility. Hence consumers could swap 2 meals (£10 worth of utility) for 1 film (£10 worth of utility) without changing their utility.
2. Since each firm sets price equal to marginal cost MC , the MC of the last meal is £5 and the MC of the last film is £10.
3. Labour earns the same wage rate in both industries in competitive equilibrium. Otherwise, workers would move to the industry offering higher wages.
4. The MC of output in either industry is the wage divided by the marginal physical product of labour MPL . Higher wages raise marginal cost, but a higher MPL means fewer extra workers are needed to make an extra unit of output.
5. Wages are equal in the two industries but the marginal cost of meals (£5) is half the marginal cost of films (£10). Hence, the MPL is twice as high in the meals industry as in the film industry.
6. Hence reducing film output by 1 unit, transferring the labour thus freed to the meals industry, raises output of meals by 2 units. The MPL is twice as high in meals as in films. Feasible resource allocation between the two industries allows society to swap 2 meals for 1 film.
7. Step 1 says that consumers can swap 2 meals for 1 film without changing their utility. Step 6 says that, by reallocating resources, producers swap an output of 2 meals for 1 film. Hence there is no feasible reallocation of resources that can make society better off. Since no Pareto gain is possible, the initial position – competitive equilibrium in both markets – is Pareto-efficient.

Notice the crucial role that prices play in this remarkable result. Prices do two things. First, they ensure that the initial position of competitive equilibrium is indeed an *equilibrium*. By balancing the quantities supplied and demanded, prices ensure that the final quantity of goods being consumed can be produced. They ensure that it is a feasible allocation.

But in *competitive* equilibrium prices perform a second role. Each consumer and each producer is a price-taker and cannot affect market prices. In our example, each consumer knows that the equilibrium price of meals is £5 and the equilibrium price of films is £10. Knowing nothing about the actions of other consumers and producers, each consumer automatically ensures that the last film purchased yields twice as much utility as the last meal purchased. Otherwise that consumer could rearrange purchases out of a given income to increase her utility.

Thus by her individual actions facing given prices, each consumer arranges that 1 film could be swapped for 2 meals with no change in utility. Similarly, every

producer, merely by setting its own marginal cost equal to the price of its output, ensures that the marginal cost of films is twice the marginal cost of meals. Thus it takes society twice as many resources to make an extra film as it does to make an extra meal. By rearranging production, transferring labour between industries, society can swap 2 meals for 1 film, exactly the trade-off that leaves consumer utility unaffected.

Thus, as if by an ‘invisible hand’, prices are guiding individual consumers and producers, each pursuing only self-interest, to an allocation of the economy’s resources that is Pareto-efficient. Nobody can be made better off without someone else becoming worse off.

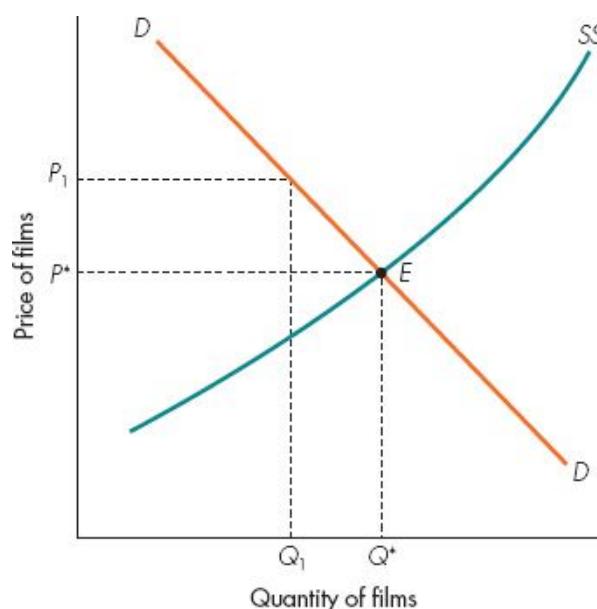


Figure 13.3 The efficient frontier

At any output such as Q_1 the last film must yield consumers P_1 pounds worth of extra utility; otherwise they would not demand Q_1 . The supply curve SS for the competitive film industry is also the marginal cost of films. If the meals industry is in competitive equilibrium, the price of a meal is also the value of its marginal utility to consumers. Thus the marginal cost of a film is not only its opportunity cost in meals but also the value of the marginal utility consumers would have derived from those meals. Hence at any film output below Q^* the marginal utility of films exceeds the marginal utility of meals sacrificed to produce an extra film. Above Q^* the marginal utility of films is less than the marginal utility of meals sacrificed. The equilibrium point E for films and the corresponding equilibrium point in the market for meals thus ensure that resources are efficiently allocated between the two industries. No reallocation could make all consumers better off.

Figure 13.3 makes the same point. DD is the market demand curve for one of the goods, say films. At a price P_1 , a quantity of films Q_1 is demanded. The last film demanded yields consumers P_1 pounds’ worth of utility; otherwise they would buy a different quantity. Hence DD shows also the marginal utility of the last unit of

films that consumers purchase. When Q_1 films are purchased, the last film yields exactly P_1 pounds' worth of extra utility to consumers.

In a competitive industry, the supply curve for films SS is also the marginal cost of films. The variable factor, labour, is paid its marginal value product in each industry. Labour mobility ensures wage rates are equal in the two industries. Hence the marginal cost of making the last film is the value of the meals sacrificed by using the last worker to make films not meals.

Prices ensure that both industries are in equilibrium. Figure 13.3 shows that, in equilibrium at E , the marginal utility of the last film equals its marginal cost. But the marginal cost of the last film is the value of meals sacrificed; the price of meals multiplied by the meals forgone by using labour to make that last film. However, the meals industry is also in equilibrium. An equivalent diagram for the meals industry shows that the equilibrium price of meals is also the marginal utility of the last meal purchased. Hence the value of meals sacrificed to make the last film is also the marginal utility of the last meal multiplied by the number of meals sacrificed.

Thus, provided the *meals* industry is in competitive equilibrium, the marginal cost curve for the *film* industry is the extra pounds' worth of utility sacrificed by using scarce resources to make another film instead of extra meals. It is the opportunity cost in utility terms of the resources being used in the film industry. And equilibrium in the film industry, by equating the marginal utility of films to the marginal utility of the meals sacrificed to make the last film, guarantees that society's resources are allocated efficiently.

At any output of films below the equilibrium quantity Q^* , the marginal consumer benefit of another film exceeds the marginal consumer valuation of the meals that would have to be sacrificed to produce that extra film. At any output of films above Q^* , society is devoting too many resources to the film industry. The marginal value of the last film is less than the marginal value of the meals that could have been produced by transferring resources to the meals industry. Competitive equilibrium ensures that there is no resource transfer between industries that would make all consumers better off.

CONCEPT 13.1

GENERAL VS PARTIAL EQUILIBRIUM: AN EXAMPLE FROM SCHOOL POLICY

In discussing the efficiency properties of a competitive equilibrium we have implicitly followed a *general equilibrium* approach. By that, we mean a situation whereby multiple markets are simultaneously in equilibrium. For example, in the analysis in Section 13.2 we have considered two markets for final goods (films and meals) and one market for inputs (the labour market). In general equilibrium we analyse how different markets are linked to and interact with each other.

This is a different approach from the one we have used in previous chapters, where we have focused mainly on what happens in a single market. When we analyse just a single market, without looking at any interaction with other markets in the economy (remember the expression ‘other things equal’ that we have used widely in previous chapters), we adopt a *partial equilibrium* approach.

Whatever approach is more suitable in analysing a particular case depends on the objective of the analysis itself. In many cases, a partial equilibrium analysis of a particular market is fine if the objective is to understand that particular market only.

When we are interested in analysing how different markets are linked together, a general equilibrium approach is preferred. The differences between a partial equilibrium and a general equilibrium approach are more evident when we evaluate government policies. Here is an example on school policy.

Heckman¹ et al. (1998) studied the partial and general equilibrium effects of a particular school policy: a \$500 tuition subsidy to college students. The partial equilibrium effect will focus on the effect that such a policy has on the college students, everything else constant. They found that a \$500 tuition subsidy leads to an increase of 5.3 per cent in college attendance. This is quite intuitive; with such a subsidy college fees become less expensive and more students can go to college.

However, this is the partial equilibrium effect only.

To get the general equilibrium we need to understand how the effect of the policy is linked to other markets. In particular, Heckman et al. focused on the labour market for college graduates. Now there are two markets linked together: the market for colleges and the labour market for college graduates.

They found that, once we take into account the link between the two markets, the result of the policy is an increase in college students of only 0.49 per cent. Why is that?

In response to the tuition subsidy more people go to college. This makes high school graduates more scarce in the labour market and college graduates more common. As a result, wages of college graduates will fall (higher labour supply of college graduates in the labour market), while wages of high school graduates will increase. Rational students will anticipate this effect and so the result of the policy will be mitigated.

Source: J. Heckman, L. Lochner and C. Taber, ‘General equilibrium treatment effects: A study of tuition policy’, *American Economic Review* 88, no. 2 (1998): 381–386. © 1998 James J. Heckman, Lance Lochner and Christopher Taber.

Equity and efficiency

We showed that under certain conditions – that is, when all markets are perfectly competitive – an economy can attain a particular Pareto-efficient allocation. This result is known as the *first theorem of welfare economics*. The previous section showed however that there are many possible Pareto-efficient allocations, each with a different distribution of utility between different members of society. What determines each one?

People have different innate abilities, human capital and wealth. These differences mean people earn different incomes in a market economy. They also affect the pattern of consumer demand. Brazil, with a very unequal distribution of income and wealth, has a high demand for luxuries such as servants. In more egalitarian Denmark, nobody can afford servants.

Different inheritances of ability, capital and wealth thus imply different demand curves and determine different equilibrium prices and quantities. In principle, by varying the distribution of initial income-earning potential, we could make the economy pick out each possible Pareto-efficient allocation as its competitive equilibrium. This result is known as the *second theorem of welfare economics*.

The second welfare theorem implies that a government can have a role in determining which efficient allocation is decided by the market. The government is elected to express the value judgements of the majority. If the market gets the economy to the Pareto-efficient frontier, the government can make the value judgement about which point on this frontier the economy should attain. Different efficient allocations correspond to different initial distributions of income-earning potential in a competitive economy. The government could redistribute income and wealth through taxation and welfare benefits in such a way that the market will then attain a particular efficient allocation.

This seems a powerful case for the free enterprise ideal. The government should let markets get on with the job of allocating resources efficiently. We do not need regulations, investigatory bodies or state-run enterprises. The government can redistribute income without impairing the efficient functioning of a free market economy. The right-wing case can be backed up by rigorous economic arguments.

However, the left-wing case can also be made. *Under certain conditions* free markets lead to a Pareto-efficient allocation. The right believes that those conditions tend to hold in reality and, even if they don't, that situation does not seriously challenge the case for a free market economy. The left believes that the fact that those conditions may not hold in reality is so serious that substantial government intervention is necessary to *improve* the way the economy works.

CONCEPT 13.2

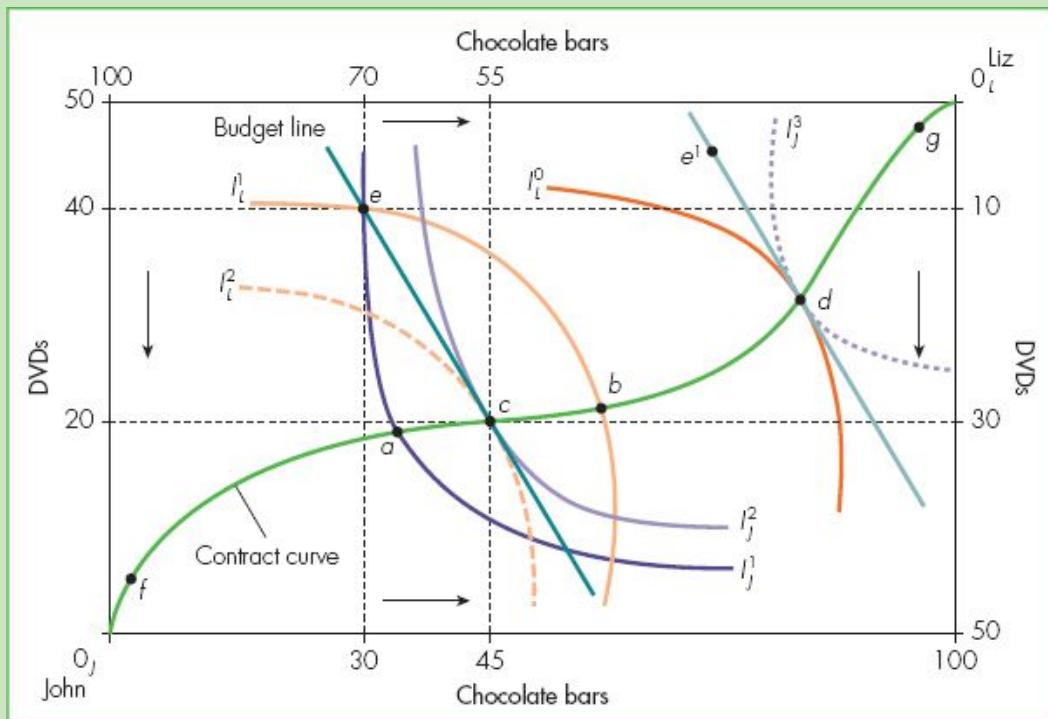
EQUITY VS EFFICIENCY IN TRADING: THE EDGEWORTH BOX

Here we show how free trading between agents can be Pareto efficient.

Suppose an economy has two agents, John and Liz. There are two goods in this economy, chocolate bars and DVDs. Each agent is born with an amount of both goods. Suppose Liz is born with 70 chocolate bars and 10 DVDs. This is called Liz's *endowment*. John instead has an initial endowment of 30 chocolate bars and 40 DVDs. The total available amount of chocolate bars is therefore 100 (5 70 1 30) and we have 50 DVDs available. There is no production activity. John and Liz must decide between consuming their initial endowment or trading with each other.

Even though this example is very simplistic and stylized, we can still use it to gather general insights on how trading can achieve a Pareto-efficient allocation of goods and how equitable can be that allocation.

To analyse this simple economy, we use a graphical device known as the Edgeworth box (named after the British economist, Francis Ysidro Edgeworth). This is shown in the figure below.



To construct the Edgeworth box, we take two graphs (such as that in Figure 5.3), one for each agent, and we rotate one of the graphs and put it on top of the other to form a box. The length and height of the Edgeworth box are given by the total amount of goods available in the economy. In our case, the length is given by 100 chocolate bars and the height by 50 DVDs. Every point inside the box denotes a particular allocation of goods between the two individuals. Point e represents the initial endowments of Liz and John. At that point, Liz has 70 chocolate bars and 10 DVDs while John has 30 chocolate bars and 40 DVDs.

For both individuals, we draw standard indifference curves representing their preferences for chocolate bars and DVDs. The origin of the graph representing John's indifference curves is the bottom-left corner. His indifference curves have the usual shape. In the graph we plot three of them, I_J^1 , I_J^2 and I_J^3 . Indifference curves further away from the origin denote higher utility. Therefore John's utility increases as we move from the bottom-left corner towards the top-right corner. For Liz's indifference curves, the opposite is true. Liz's utility increases as we move from the top-right of the box towards the bottom-left. We display three possible indifference curves for Liz, I_L^1 , I_L^2 and I_L^3 .

Finally, we assume that there is a third agent, Robert. He does not have any goods and he does not trade. He does just one thing: he shouts relative prices to John and Liz. For example, making p the price of chocolate bars and q the price of DVDs, he can shout $p/q = 2$ or $p/q = 0.5$, and so on. Robert is called an auctioneer. John and Liz take those prices as given and every time they hear a relative price, they decide how much of both goods to trade with each other. Trade occurs only if it is mutually beneficial.

Of course, John and Liz can keep and consume their initial endowment, meaning they do not need to engage in trading. The question is: if they trade, can they achieve a better allocation compared to their initial endowment?

By looking at the Edgeworth box, we can see that the initial allocation is point e . That point lies on Liz's indifference curve I_L^1 and on John's indifference curve I_J^1 . They both benefit from trading. They can trade with each other to reach any allocation on the curve connecting points a and b . Any such allocation is a Pareto gain compared to their initial endowment allocation. Which allocation between a and b will be chosen? This depends on the relative prices shouted by Robert. In the box we depict a budget line. This represents a line with a slope given by the negative of the relative price of chocolate bars and DVDs; that is, $-2p/q$. If the relative price shouted by Robert is the one that gives rise to the budget line depicted in the box, then the chosen allocation is point c , whereby John trades 20 DVDs with Liz in exchange for 15 chocolate bars. At point c , John ends up with 20 DVDs and 45 chocolate bars, while Liz ends up with 30 DVDs and 55 chocolate bars. Once allocation c is reached, no further gain from trading can be achieved.

At allocation c , we have that Liz's marginal rate of substitution (MRSL) is equal to John's marginal rate of substitution (MRSJ). Liz's indifference curve I_L^2 is tangent to John's indifference curve I_J^2 and they are both tangent to the budget line. Therefore at point c we have $MRSL = MRSJ = p/q$ in absolute values. Allocation c is Pareto-efficient. There is no way to modify the allocation of goods between the two agents without decreasing the utility of at least one of them. Allocation c represents a first-best allocation.

Notice the important result here: there is a relative price level that sustains a Pareto-efficient allocation of goods. This is the message of the first welfare

theorem. Competitive markets, through the price system, allocate resources in the most efficient way.

The set of all feasible Pareto-efficient allocations is called the contract curve. At each point on the contract curve it must be true that the marginal rate of substitution of the two agents is equal.

Let's now turn to the issue of equity. Any allocation on the contract curve is Pareto-efficient. But allocations such as f or g imply that one of the agents obtains almost all goods while the other gets almost nothing. They are not very equitable. The chosen allocation on the contract curve depends on the initial endowment. Suppose the initial endowment is point e1. John is born wealthier than Liz. By trading, Liz and John can now reach allocation d on the contract curve. Suppose that Robert now plays another role, that of a government. Robert feels that allocation d is not equitable because it results in John having most of both goods. He would prefer John and Liz to share a more even number of both goods as a result of their trading, like in allocation c. He can do the following: before trading takes place he imposes lump sum taxes and transfers to Liz and John.² In particular, he taxes John by taking some of his chocolate bars and some of his DVDs and transfers them to Liz. If he does a good job, he changes the initial endowments of John and Liz from e1 to e. Then he lets John and Liz trade so that they can achieve allocation c.

This result shows that, by properly redistributing the initial endowments of the agents, we may sustain a Pareto-efficient allocation of resources. This is the main message of the second welfare theorem.

13.3 Distortions and the second-best

Competitive equilibrium is efficient because the independent actions of producers setting marginal cost equal to price, and consumers setting marginal benefits equal to price, ensure that the marginal cost of producing a good just equals its marginal benefit to consumers.

Taxation as a distortion

To finance subsidies to the poor, a government must tax the incomes of rich people or the goods rich people buy. Suppose everyone buys meals, but only the rich can afford to go to the cinema. A subsidy for the poor can be financed by a tax on films.

In Section 4.9 we showed that the introduction of a tax in a market has the effect of creating a wedge between the price paid by the consumers and the price received by the suppliers. Figure 13.4 shows the effects of introducing a tax on the market for films. The pre-tax price of films paid by consumers (P_1) exceeds the post-tax price received by makers of films (P_2). The difference between the two prices is the tax

on each film. Consumers equate the tax-inclusive price to the value of the marginal benefit they receive from the last film, but suppliers equate the marginal cost of films to the lower net-of-tax price of films.

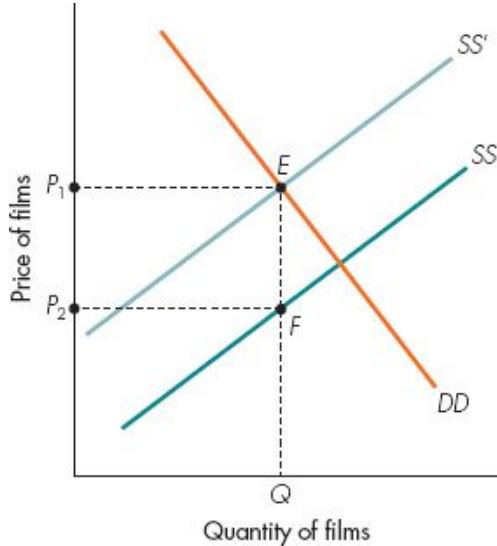


Figure 13.4 A tax on films

DD shows the demand for films and the marginal benefit of the last film to consumers. SS shows the quantity of films supplied at each price received by producers and is also the marginal social cost of producing films. Suppose each unit of films bears a tax equal to the vertical distance EF. To show the tax-inclusive price required to induce producers to produce each output, we must draw the new supply curve SS' that is a constant vertical distance EF above SS. The equilibrium quantity of films is Q. Consumers pay a price P_1 , producers receive a price P_2 and the tax per film is the distance EF. At the equilibrium quantity Q the marginal benefit is P_1 but the marginal social cost is P_2 . Society would make a net gain by producing more films. Hence the equilibrium quantity.

Because of the tax the price system no longer equates the marginal cost of making films with the marginal benefit of consuming films. The marginal benefit of another film exceeds its marginal cost. The tax on films induces too few films compared to what society would like.

Earlier, we showed that the marginal cost of a film equals the value of the extra meals that society could have had instead. When films are taxed, the marginal benefit of another film exceeds its marginal cost, and hence exceeds the marginal benefit of the extra meals that society could have had by using resources differently. By transferring labour from meals into films, society could make some people better off without making anyone else worse off.

A **distortion** exists if society's marginal cost of producing a good does not equal society's marginal benefit from consuming that good.

A similar argument holds for any other commodity we tax. A tax causes a discrepancy between the price the purchaser pays and the price the seller receives. The ‘invisible hand’ no longer equates marginal benefits of resources in different uses.

The choice between efficiency and equity is now clear. If the economy is perfectly competitive, and if the government is happy with the current income distribution, competitive free market equilibrium is efficient and the income distribution desirable.

However if, as a pure value judgement, the government dislikes this income distribution, it has to tax some people to provide subsidies for others. Yet the very act of raising taxes introduces a **distortion**. The resulting equilibrium has a more desirable distribution but is less efficient. Governments may have to make trade-offs between efficiency and equity.

One explanation for differing political attitudes to the market economy is a difference in value judgements about equity. Later, we will see that there may also be disagreements in positive economics. We consider other distortions in the next section. Before leaving our tax example, there is one final point to make.

The second-best

The **first-best allocation** has no distortions and is fully efficient.

When there is no distortion in the market for *meals*, a tax on *films* leads to an inefficient allocation. If we could abolish the tax on films, neither industry would be distorted and we get the **first-best allocation**.

Suppose, however, that we cannot get rid of the tax on films. The government needs tax revenue to pay for national defence or its EU budget contribution. Given an unavoidable tax on films, at least it should not tax meals as well.

This plausible view is in fact *quite wrong*. Suppose both industries are in equilibrium but there is a tax on films. Above, we saw that too few films are produced and consumed. By implication, too many meals are therefore produced and consumed. Given an inevitable tax on films, a tax on meals would help not hinder.

A suitable tax on meals could restore the original relative price of meals and films. With only two goods, this would restore the first-best. However, there is always a third good – leisure. Households reduce consumption of leisure in order to supply labour for work. Taxing meals and films achieves the right balance between meals and films, but makes the price of both wrong relative to the price of leisure. With higher taxes, the net wage falls, changing the implicit price of leisure. Therefore,

even if we can offset the distortion in the film market by introducing an appropriate tax in the meal market, the first-best allocation cannot be attained.

The **second-best** is the most efficient outcome that can be achieved conditional on being unable to remove some distortions.

However we can achieve what is known as the *second-best* allocation. The **second-best** theory says that, if there must be a distortion, it is a mistake to concentrate the distortion in one market. It is more efficient to spread its effect more thinly over a wide range of markets.

Several applications of this general principle are found in the ensuing chapters. The real world in which we live provides several inevitable distortions. Given their existence, the argument of this section implies that the government may *increase* the overall efficiency of the whole economy by introducing *new* distortions to offset those that already exist. By now you will want to know the source of these inevitable distortions that the government could take action to offset.

13.4 Market failure

In the absence of any distortions, competitive equilibrium is efficient. We use the term *market failure* to cover all the circumstances in which market equilibrium is inefficient. Distortions then prevent the ‘invisible hand’ from allocating resources efficiently. We now list the possible sources of distortions that lead to market failure.

Imperfect competition

Only perfect competition makes firms equate marginal cost to price and thus to marginal consumer benefit. Under imperfect competition, producers set a price above the marginal cost. Since consumers equate price to marginal benefit, marginal benefit exceeds marginal cost in imperfectly competitive industries. Such industries produce too little compared to the efficient level. Increasing the level of competition in an imperfectly competitive market would result in higher output produced. This would add more to consumer benefit than to production costs (or the opportunity cost) of the resources used.

Equity, taxation and public goods

Redistributive taxation induces allocative distortions by driving a wedge between the price the consumer pays and the price the producer receives. So far, when discussing the goods produced in the market, we have considered private goods.

Private goods are those that can be consumed only by the buyer. For example, if you buy a can of Coke, you pay for it and you drink it. Other consumers cannot drink the same can of Coke. In contrast to private goods, public goods are those that, if consumed by one person, can be consumed by others in exactly the same quantity. National defence is an example. Since you get the same quantity of national defence as everyone else, *whether or not you pay for it*, you never buy national defence in a private market. Therefore, in the case of public goods, we have goods that society would like to consume but the private market mechanism cannot provide (or will under-provide). Taxes and public goods are analysed in detail in Chapter 14.

Externalities

Externalities arise when one person's actions have direct costs or benefits for other people, but the individual does not take these into account. Externalities are things like pollution, noise and congestion. Much of the rest of this chapter examines this distortion. The problem arises because there is no market for things like noise. Hence markets and prices cannot ensure that the marginal benefit you get from making a noise equals the marginal cost of that noise to other people.

Asymmetric information

In Chapter 12 we saw how moral hazard and adverse selection inhibit the setting-up of insurance markets to deal with risk. The fact that there exists imperfect information in certain markets may lead to a failure in such markets.

Under asymmetric information, one party in a market transaction has more information than the other party. For example, a seller may know the true quality of the good she is selling while the buyer does not.

Suppose that buyers want to buy used cars. There are various sellers in the market, some selling high-quality used cars, some selling low-quality used cars. If the buyers cannot tell the difference between low and high quality, they will probably be unwilling to pay much for a used car (they always face the possibility of getting a low-quality used car). As a result, the sellers with high-quality used cars may end up selling them at a price that is lower than their value, meaning it is unprofitable for those sellers to stay in the market. In practice, under asymmetric information, the existence of the low-quality product drives the high-quality product out of the market. This is a market failure since the market for the high-quality product is eliminated even if the buyers value it at more than the cost of producing it. To mitigate the effects of asymmetric information, sellers can use warranties. By issuing longer warranties, sellers can signal that their cars are high quality. The case of used cars is known as the market for 'lemons' (whereby lemons means low-quality used cars) – a typical example of adverse selection.

13.5

Externalities

An **externality** is a cost or benefit related to the production or consumption of some good that is imposed on others in a way other than by charging prices.

Externalities arise when someone engages in production or consumption activities that affect others but none pay or receive compensation for those effects.

A chemical firm discharges waste into a lake, polluting the water. It affects the production of anglers (fewer fish, harder to catch) or the consumption of swimmers (dirty water). Without a ‘market’ for pollution, the firm can pollute the lake without cost. It ignores the cost that pollution imposed on anglers and swimmers. This is an example of a negative production externality.

Conversely, by painting your house you make the whole street look nicer and give consumption benefits to your neighbour. However, in deciding how much to paint you do not take into account the benefits you provide for your neighbour. This is an example of a positive consumption externality.

Externalities are side effects of production and consumption. They have economic value (positive or negative) but there is no market for them. When an externality is present in a market there will be a divergence between the individual’s private marginal costs and benefits and society’s marginal costs and benefits.

In a competitive market without externalities and distortions, the marginal private benefits coincide with the social marginal benefits and the marginal private costs coincide with the marginal social costs. Therefore, in a competitive equilibrium an efficient allocation implies that marginal social benefits equal marginal social costs. When an externality is present, this result does not hold and an inefficient market allocation is created.

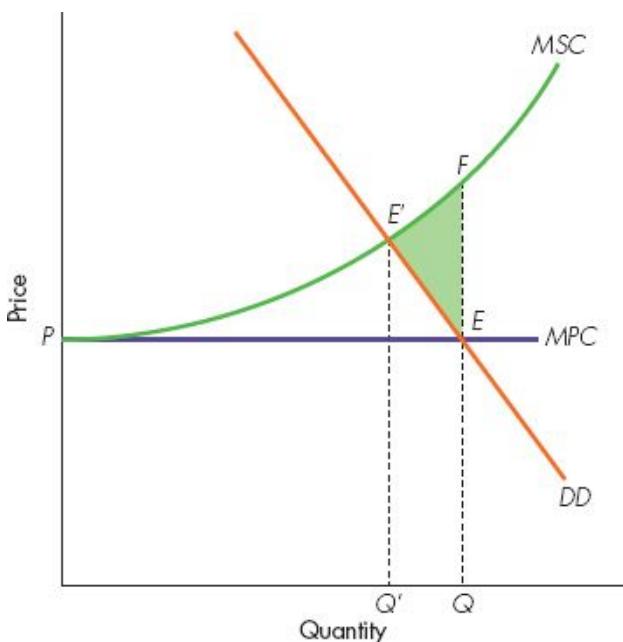


Figure 13.5 The social cost of a negative production externality

Competitive equilibrium occurs at E. The market clears at a price P, which producers equate to marginal private cost MPC. But pollution causes a production externality which makes the marginal social cost MSC exceed the marginal private cost. The socially efficient output is at E', where marginal social cost and marginal social benefit are equal. The demand curve DD measures the marginal social benefit because consumers equate the value of the marginal utility of the last unit to the price. By inducing an output Q in excess of the efficient output Q' free market equilibrium leads to a social cost equal to the area E'FE. This shows the excess of social cost over social benefit in moving from Q' to Q.

Divergences between private and social costs and benefits

Suppose a chemical firm pollutes a river, the quantity of pollution rising with output. Downstream, companies use river water as an input in making sauce for baked beans. The production of chemicals creates a negative externality for the downstream companies.

Figure 13.5 shows the marginal private cost *MPC* of producing chemicals. For simplicity, we treat *MPC* as constant.³ It also shows the marginal *social cost MSC* of chemical production. At any output, the divergence between marginal private cost and marginal social cost is the marginal *production externality*.

A production externality makes private and social marginal costs diverge. The demand curve *DD* shows how much consumers will pay for the output of the chemical producer. If that firm is a price-taker, equilibrium is at E and the chemical producer's output is Q, at which the marginal private cost equals the price of the firm's output.

At this output Q, the marginal social cost *MSC* exceeds the marginal social benefit of chemicals, given by the height of the demand curve *DD*. The market for

chemicals ignores the production externality inflicted on other firms. At Q , the marginal social benefit of the last output unit is less than the marginal social cost inclusive of the production externality. Output Q is inefficient. By reducing the output of chemicals, society saves more in social cost than it loses in social benefit. Society could make some people better off without making anyone worse off.

The efficient output is Q_9 , at which the marginal social benefit equals the marginal social cost. E_9 is the efficient point. How much does society lose by producing at the free market equilibrium E and not the efficient point E' ? The vertical distance between the marginal social cost MSC and the marginal social benefit shows the marginal social loss of producing the last output unit. By over-expanding from Q' to Q , society loses the area E_9FE in Figure 13.5. This is the social cost of the market failure caused by the production externality of pollution.⁴

Production externalities make social and private marginal costs diverge. *A consumption externality makes private and social marginal benefits diverge*. Figure 13.6 shows a positive consumption externality. Planting roses in your front garden also makes your neighbours happy.

With no production externality, MPC is both the private and social marginal cost of planting roses. It is the cost of the plants and the opportunity cost of your time. DD is the marginal private benefit. Comparing your own costs and benefits, you plant a quantity Q of roses.

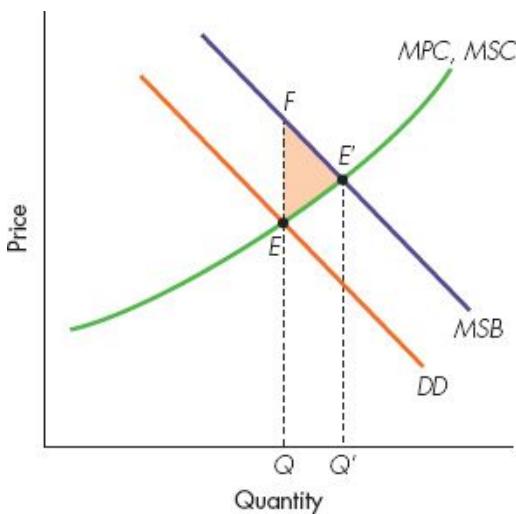


Figure 13.6 A beneficial consumption externality

With no production externality, marginal private cost and marginal social cost coincide. DD measures the marginal private benefit, and the market equilibrium occurs at E . The beneficial consumption externality makes marginal social benefit MSB exceed marginal private benefit. E' is the socially efficient point. By producing Q instead of the efficient output Q' , free market equilibrium wastes the area $EE'E'$.

But you ignore the consumption benefit to your neighbours. The marginal social benefit MSB exceeds your marginal private benefit. The free market equilibrium is

at E , but the efficient output is Q' since marginal social benefit and marginal social cost are equated at E .

Society could gain the area EFE' , the excess of social benefits over social costs, by increasing the quantity of roses from Q to Q' . This triangle measures the social cost of the market failure that makes equilibrium output too low.

From our analysis we can draw a general conclusion: free markets will lead to over-production (*under-production*) of goods with negative (*positive*) externalities.

CASE 13.1

EXTERNALITIES AND THE LONDON 2012 OLYMPIC GAMES

The summer Olympic Games is among the most important sporting events in the world. Moreover, such events provide the possibility for environmental and economic improvements in the host city. In 2012 London hosted the Olympic Games.

There are many costs involved in preparing for such a massive sporting event and there are also many benefits accruing from it. Some of those costs and benefits can be easily measured in monetary terms. For example, according to recent estimates the construction of the venues for the London Olympic Games contributed £5.8 billion to UK GDP in 2012. The boost in tourism was estimated to contribute £2 billion to GDP. We can measure such a benefit because we have a market for tourism. However, many other benefits and costs associated with the Olympic Games are intangible.

By intangible, we mean costs and benefits that will impact the well-being of many people but that will not in general have market prices. In other words, intangible benefits are positive externalities and intangible costs are negative externalities. The Olympic Games is a source of externalities.

For example, hosting the Olympic Games normally boosts the national pride of the hosting country. This can be seen as a positive externality (a sort of 'feelgood' effect) but there is no market for national pride. Another possible positive externality that may be produced is related to environmental improvements through creation of new green spaces and recreational areas. The Olympic Games will probably create a sport and cultural legacy in the UK. This will enhance and accelerate investment in sporting facilities not only within the Olympic zone (and the areas immediately surrounding it) but also in other parts of the UK. It may contribute to increased participation in sport, and this will be expected to promote healthy living.

Cultural and social events may also improve during and after the Olympic Games.

Obviously there are also negative externalities. In some parts of London traffic congestion was a big problem during the Games. Construction of the Olympic

site caused disruption for local residents. As a result of building all the facilities, pollution may have increased in the Olympic zone.

All those intangible benefits and costs are difficult to measure in monetary terms. In 2005 the UK Department for Culture, Media and Sport commissioned PricewaterhouseCoopers to study the possible costs and benefits of the London 2012 Games. An interesting finding of that study is the estimated willingness of London households to pay for the intangible benefits of the Olympic Games. On average, London households are willing to pay £22 each per annum for ten years in order to host the Olympic Games. Therefore, there is a positive valuation of the intangible benefits that London 2012 can bring to Londoners.



Wenlock and Mandeville, the London 2012 Mascots. © Richard Allen/Alamy

Sources: PricewaterhouseCoopers, *Olympic Games impact study: Final report*, December 2005; Oxford Economics, 'The economic impact of London 2012', July 2012.

Property rights and externalities

Your neighbour's tree obscures your light – a negative consumption externality. If the law says that you must be compensated for any damage suffered, your neighbour has to pay up or cut back the tree.

She likes the tree and wants to know how much it would take to compensate you to leave it at its current size. Figure 13.7 shows the marginal benefit MB that she gets from the last inch of tree and the marginal cost MC to you of that last inch. At the tree's current size S_1 , the total cost to you is the area $OABS_1$. This is the marginal cost OA of the first inch, plus the marginal cost of the second inch, and so on to the existing size S_1 . The area $OABS_1$ is what you need in compensation if the tree size is S_1 .

Your neighbour is about to pay up when her daughter, an economics student, points out that, at size S_1 , the marginal benefit of the last inch to her is less than the marginal cost to you, the amount you must be compensated for that last inch of the tree. It is not worth her mother having a tree this big. Nor, she points out, is it worth cutting the tree down altogether. The first inch yields a higher marginal benefit to her than the amount that you need in compensation to offset your marginal cost of that first inch. A tiny tree has little effect on your light.

At the efficient tree size S^* , the marginal benefit to your neighbour equals the marginal cost to you. Above S^* , she cuts back the tree, since the marginal cost (and compensation) exceeds her marginal benefit. Below S^* , she increases the tree size, and pays you marginal compensation that is less than her marginal benefit. At the efficient size S^* , your total cost is the area $OAES^*$. This is the compensation you are paid. Property rights are the power of residual control, including the right to be compensated for externalities.

Since a larger tree benefits one party but hurts the other, *the efficient tree size, and therefore the efficient quantity of the externality, is not zero*. It is where the marginal benefit equals the marginal cost.

Property rights are the power of residual control, including the right to be compensated for externalities.

Property rights affect who compensates whom, a distributional implication. Suppose there is no law requiring compensation. Instead of letting her tree grow to S_1 , inflicting a huge cost on you, you bribe your neighbour to cut it back. You compensate her for the loss of her marginal benefit. You would pay to have the tree cut back as far as S^* but no further. Beyond that size, you pay more in compensation for loss of marginal benefit than you save yourself in lower cost of the externality. So you pay a *total* of S^*EDS_1 to compensate for the loss of benefit in cutting the tree back from S_1 to S^* . Who has the property rights determines who pays whom, but does not affect the efficient quantity that the bargain determines. It is always worth reaching the point at which the marginal benefit to one of you equals the marginal cost to the other.

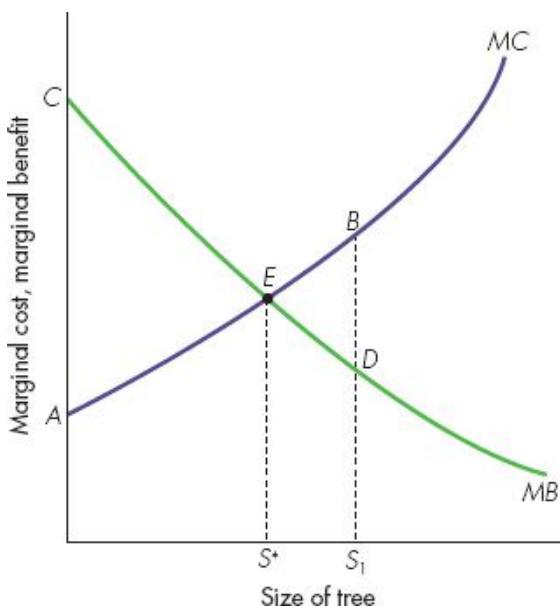


Figure 13.7 The efficient quantity of an externality

MB and MC measure the marginal benefit to your neighbour and marginal cost to you of a tree of size S . The efficient size is S^* , where the marginal cost and benefit are equal. Beginning from a size S_1 , you might bribe your neighbour the value S^*EDS_1 , to cut back to S^* . Below S^* you would have to pay more than it is worth to you to have the tree cut back further. Alternatively, your neighbour might pay you the value $OAES^*$ to have a tree of size S^* . Property rights, in this case whether you are legally entitled to compensation for loss of light to your garden, determine who compensates whom but not the outcome S^* of the bargain.

Property rights have a distributional implication – who compensates whom – but also achieve the efficient allocation. They set up the ‘missing market’ for the externality. The market ensures that the price equals the marginal benefit and the marginal cost, and hence equates the two.

Economists say that property rights ‘internalize’ the externality. The relationship between property rights, efficiency and externalities is known as the *Coase theorem*.⁵

This theorem says that, when there are no transaction costs and trading externalities is possible, then the trading mechanism will lead to an efficient outcome independent of the initial allocation of the property rights. For example, consider two firms: one is polluting and doing so negatively affects the other firm. In this case, it does not matter if we assign the right to pollute to the polluting firm or, alternatively, the right not to be polluted to the other firm. Once the property rights are assigned, the externality will be internalized.

The basic idea behind the Coase theorem is: if people must pay for it they will take its effects into account in making private decisions and there will no longer be market failure. Why, then, do externalities, like congestion and pollution, remain a problem? Why don’t private individuals establish the missing market through a

system of bribes or compensation? A free-rider, unable to be excluded from consuming a good, has no incentive to buy it.

There are two reasons why it is hard to set up this market. The first is the cost of organizing the market. A factory chimney dumps smoke on a thousand gardens nearby, but it is costly to collect £1 from each household to bribe the factory to cut back to the efficient amount. Second, there is a **free-rider** problem.⁶

Someone knocks on your door and says: ‘I’m collecting money from people who mind the factory smoke falling on their gardens. The money will be used to bribe the factory to cut back. Do you wish to contribute? I am going round 1000 houses nearby.’ Whether you mind or not, you probably say: ‘I don’t mind, and won’t contribute.’ If everybody else pays, the factory will cut back and you cannot be prevented from getting the benefits. The smoke will not fall exclusively on your garden just because you alone did not pay. Regardless of what other people contribute, your dominant strategy is to be a free-rider. Everyone else reasons similarly; hence no one pays, even though you are all better off paying and getting the smoke cut back.

MATHS 13.1

INTERNALIZING A NEGATIVE EXTERNALITY USING PROPERTY RIGHTS

Consider a firm that is polluting a lake in order to produce. This is called firm A. There is another firm, B, which uses the fish in the lake. All markets for outputs are competitive. The total cost function of the polluting firm (firm A) is:

$$TC_A = TC_A(Q_A, P_A)$$

That is a function of the quantity produced (Q_A) and the level of pollution (P_A).

We assume that the total cost of firm A is increasing with the output produced:

$$\frac{\partial TC_A}{\partial Q_A} > 0$$

This means that if we increase by a small amount (δQ_A) the quantity produced, the total cost increases. We assume that the total cost of firm A is decreasing with the pollution level $\frac{\partial TC_A}{\partial P_A} \leq 0$.

More pollution implies lower costs for the firm. Think about the case in which, to pollute less, the firm must invest in an expensive cleaner technology. If p is the market price of output for firm A, then the profit function of that firm is $\pi_A = PQ_A - (Q_A, P_A)$.

Firm A chooses the optimal quantity to produce (Q_A) at which the marginal revenue is equal to the marginal cost of producing that quantity: $p = \frac{\partial TC_A}{\partial Q_A}$.

Similarly, the optimal quantity of pollution that maximizes profits is where marginal revenue of pollution (in this case, zero) is equal to the marginal cost of pollution:

$$-\frac{\partial TC_A}{\partial P_A} = 0$$

The firm chooses a level of pollution such that the cost of an extra unit of pollution is zero. Since the higher is pollution, the lower is the total cost of the firm, we should expect that the level of pollution that solves that condition to be quite high.

Firm B has the total cost function $TC_B = TC_B(Q_B, P_A)$, with the properties $\frac{\partial TC_B}{\partial Q_B} > 0$ and $\frac{\partial TC_B}{\partial P_A} > 0$.

This means that the total cost of firm B increases with output produced and with the pollution made by firm A. The externality problem is the following: firm A, in deciding how much to pollute, does not take into account the effects that its decision has on firm B. Make f the market price of fish for firm B. The profit of firm B is then given by $\pi_B = fQ_B - TC_B(Q_B, P_A)$.

Suppose we give the right to pollute to firm A. Firm A can sell its right to firm B. The profit function of firm A becomes $\pi_A = pQ_A - TC_A(Q_A, P_A) + qP_A$, where q is now the price that firm A can get by selling its right to pollute to firm B. For firm B, the profit function is now $\pi_B = fQ_B - TC_B(Q_B, P_A) - qP_A$. For firm A, the optimal level of pollution that maximizes profits is given by the condition:

$$-\frac{\partial TC_A}{\partial P_A} + q = 0 \Rightarrow \frac{\partial TC_A}{\partial P_A} = q$$

That condition simply says marginal cost is equal to marginal revenue from polluting (now equal to q).

For firm B, the quantity of pollution that maximizes its profit is given by the condition:

$$-\frac{\partial TC_B}{\partial P_B} - q = 0 \Rightarrow \frac{\partial TC_B}{\partial P_B} = q$$

Equations (1) and (2) imply that the price q should satisfy the following: $-\frac{\partial TC_A}{\partial P_A} = \frac{\partial TC_B}{\partial P_B}$.

In deciding the optimal level of P_A , firm A now takes into account the effect that its decision has on firm B. In particular, it must set a level of pollution such that the marginal private cost of polluting ($-\frac{\partial TC_A}{\partial P_A}$) is equal to the marginal social cost of polluting ($-\frac{\partial TC_B}{\partial P_B}$). So, by assigning the property rights we can obtain the efficient level of pollution. You can try to work out the case

in which firm B has the right not to be polluted and can sell this right to firm A; does the result above still hold?

13.6 Environmental issues and the economics of climate change

When there is no implicit market for pollution, pollutants are overproduced. Private producers ignore the costs they impose on others. In equilibrium, social marginal cost exceeds social marginal benefit.

The most topical environmental externality we are currently facing is global warming, or climate change; that is, the rise in global temperature due to human activity. In particular, global warming is an externality in two main dimensions:

1. *An intergenerational dimension*: what humans are currently doing will affect future generations not yet born.
2. *An international dimension*: what a country does in terms of emissions will affect other countries.

In the case of a negative externality the government can induce private producers to take account of the costs inflicted on others by charging them (through taxes) for the divergence between marginal private and social costs.

Pollution taxes, especially for water pollution, are used in many countries. But most policy takes a different approach: imposing pollution standards to regulate the quantities of pollution allowed. For example, since the Clean Air Act 1956, UK governments have designated clean air zones in which certain pollutants, notably smoke caused by burning coal, are illegal. Table 13.1 shows a big fall in smoke pollution in the UK over time as a result of this policy.

Table 13.1 Smoke emission, UK (million tonnes per annum)

1958	1974	2003
2.0	0.8	0.1

Sources: *Digest of Environmental Protection and Water Statistics*; ONS, *Social Trends*.

For global warming, given its international dimension, things are more complicated because an effective policy should be agreed on and implemented by a coalition of governments.

Prices vs quantities

If free markets tend to over-pollute, society can cut pollution either by regulating the quantity of pollution or by using the price system to discourage such activities by taxing them. Is it more sensible to intervene through the tax system than to regulate quantities directly?

Many economists prefer taxes to quantity restrictions. If each firm is charged the same price or tax for a marginal unit of pollution, each firm equates the marginal cost of reducing pollution to the price of pollution. Any allocation in which different firms have different marginal costs of reducing pollution is inefficient. If firms with low marginal reduction costs contract further and firms with high marginal reduction costs contract less, lower pollution is achieved at less cost.

The main problem with using just taxes is uncertainty about the outcome. Suppose pollution beyond a critical level has disastrous consequences, for example irreversibly damaging the ozone layer. By regulating the quantity directly, society can ensure a disaster is avoided. Indirect control, through taxes or charges, runs the risk that the government does its sums wrong and sets the tax too low. Pollution is then higher than intended, and may be disastrous.

Regulating the total quantity of pollution, with spot checks on compliance by individual producers, is a simple policy that avoids the worst outcomes. However, by ignoring differences in the marginal cost of reducing pollution across different polluters, it does not reduce pollution in a way that is cost-minimizing to society.

Lessons from the United States

The US has gone furthest in trying to use property rights and the price mechanism to cut back pollution efficiently. The US Clean Air Acts established an environmental policy that includes an *emissions trading programme* and *bubble policy*.

The Acts lay down a minimum standard for air quality, and impose pollution emission controls on particular polluters. Any polluter emitting less than their specified amount gets an *emission reduction credit* (ERC), which can be sold to another polluter wanting to exceed its allocated pollution limit. Thus, the total quantity of pollution is regulated, but firms that can cheaply reduce pollution have an incentive to do so, and sell off the ERC to firms for which pollution reduction is more expensive. We get closer to the efficient solution in which the marginal cost of pollution reduction is equalized across firms.

When a firm has many factories, the bubble policy applies pollution controls to the firm as a whole. The firm can cut back most at the plants in which pollution reduction is cheapest.

Thus, the US policy combines ‘control over quantities’ for aggregate pollution, where the risks and uncertainties are greatest, with ‘control through the price system’ for allocating efficiently the way these overall targets are achieved.

The economics of climate change

There is increasing evidence that global temperatures are rising. The science of climate change means that we are also likely to see greater fluctuations in climate as well. Hence, extreme events will become much more frequent. Large parts of Bangladesh may disappear under water for ever; and English villages, from Yorkshire to Cornwall, have already experienced flash flooding. Conversely, regions of the world that are currently temperate may become arid and uninhabitable. Figure 13.8 shows the dramatic change in global temperatures in recent years.

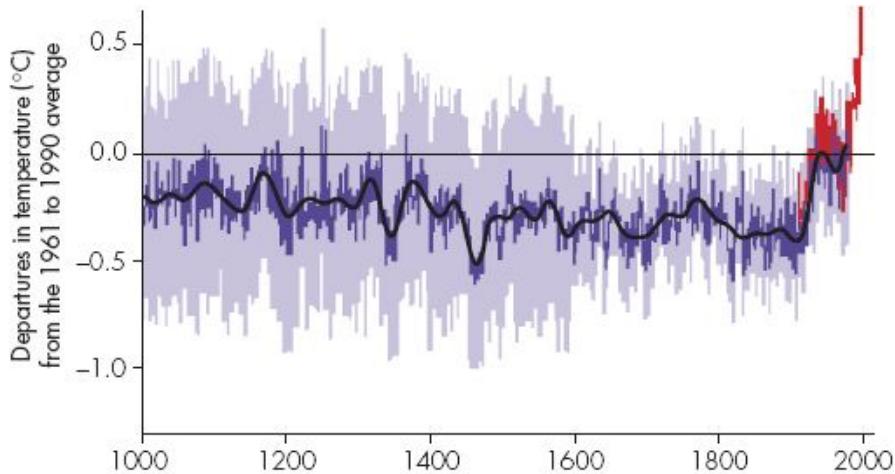


Figure 13.8 A brief history of global temperature, northern hemisphere

Data from thermometers (red) and from tree rings, corals, ice cores and historical records (blue).

Source: Intergovernmental Panel on Climate Change.

The science of climate change

The earth's climate is affected by many things, from solar radiation to the consequences of human behaviour. The ebb and flow of previous ice ages reminds us that human behaviour is not the only cause of climate change. Even so, there is increasing evidence that we must look to ourselves as a major cause of recent global warming.

In the same period in which we experienced an increase in global temperature, there was a significant increase in global CO₂ emissions due to human activity. This is shown in Figure 13.9.

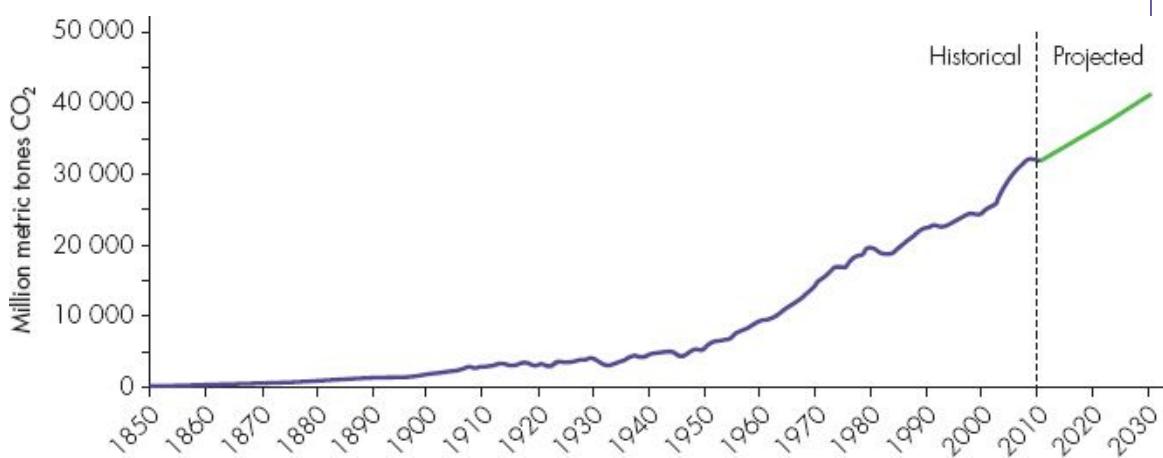


Figure 13.9 Historical evolution of global CO₂ emissions, 1850–2030

Source: Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory (2012), International Energy Agency, World Energy Outlook (2012).

Greenhouse gases – including carbon dioxide and methane – shield the earth from solar radiation, but also trap the heat underneath. Without them, all heat would escape and we would freeze to death. But we need just the right amount. Too much greenhouse gas and the earth overheats, causing global warming.

The recent build-up of greenhouse gases reflects large emissions of carbon dioxide from households, power stations and transport. This may cause ice to melt and water to expand, causing sea levels to rise. A catastrophic eventual consequence would be melting of permafrost in Siberia, releasing such volumes of methane that a large rise in temperature would then be inevitable, perhaps threatening human survival.

Carbon, a key constituent of all greenhouse gases, is a useful common denominator. Slowing, let alone reversing, global warming requires the emission of much less carbon.

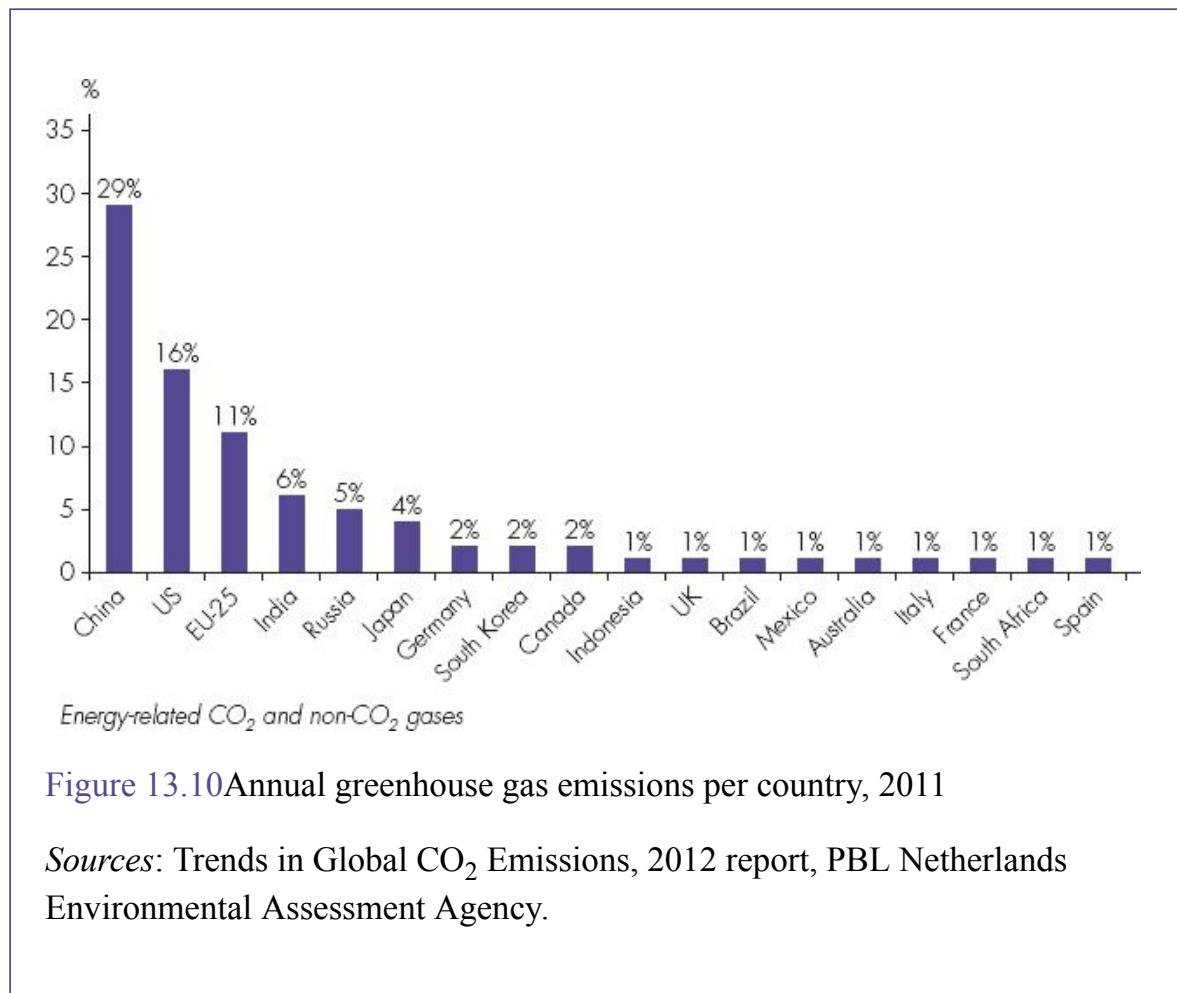
Figure 13.10 shows greenhouse gas emissions per country in 2011 (CO₂ is an important greenhouse gas but it is not the only one). As we can see, China is the country that produces the largest amount of greenhouse gases, followed by the US and then the European Union as a whole.

The Kyoto Protocol

In 1997 a group of countries signed an amendment to the UN International Treaty on Climate Change, committing themselves to cut greenhouse gas emissions. By

2006, 169 countries (though not the US) had signed. In December 2012 an agreement was reached to extend the treaty to 2020.

Developed countries accept the obligation to reduce emissions by 2012 to 5 per cent below the level of their emissions in 1990. Developing countries have not yet made a commitment but can take part in the Clean Development Mechanism. Thus, China and India ratified the protocol but are not yet bound by the commitment to reduce emissions: given their population size, rate of economic growth and future energy demands, China and India will have a huge impact on what happens with greenhouse gases.



Within the EU's overall targets, individual members can buy and sell obligations within the EU Emissions Trading Scheme (which resembles the US pollution scheme discussed earlier). The Clean Development Mechanism allows India or China to invest in emissions reduction, such as by building a cleaner power station, and sell the emissions credit to a UK or German company so that Europe then meets its overall emissions obligations.

Thus the projected total cutbacks can be achieved efficiently – those most easily able to reduce emissions cheaply do so; those for whom emissions reduction is expensive can instead purchase a credit from someone else better placed to cut back emissions cheaply.

If you think about it, this is an application of the property rights argument we have previously discussed.

Cost–benefit analysis

Even if we accept the science, what should we do, and how quickly? This gets to the core of the what, how and for whom questions of Chapter 1. The for whom question is particularly acute. How much pain should the current generation take in order to make life nicer for future generations? Can we expect China and India to slow their economic development to make life nicer for citizens in Europe and the US, who begin with many more economic advantages?

The Kyoto targets are modest, and as yet fail to include the key economies of the US, China and India, on whom much will actually depend. Kyoto supporters see these targets as the thin end of the wedge, creating a political dynamic that will create tougher targets soon; which is precisely why they are opposed by those who would potentially lose out.

In 2006 the UK government published a report on the economics of climate change written by Sir Nicholas Stern, a London School of Economics professor, and ex-Chief Economist of both the World Bank and the European Bank for Reconstruction and Development. The Stern Review (details of which are available at www.hm-treasury.gov.uk) concluded that 1 per cent of global GDP must be invested from now on if we are to head off the worst effects of climate change; and that failure to act now risks a future cost of up to 20 per cent of global GDP.

Many of the world's leading economists – including economics Nobel Prize winners Sir James Mirrlees, Amartya Sen, Joe Stiglitz and Bob Solow, and Professor Jeffrey Sachs, Director of the Earth Institute at Columbia University in New York – have come out strongly in support of the Stern Review. The principal point of subsequent debate has been the appropriate interest rate at which to discount future costs and benefits, a topic we discuss in Activity 13.1. The decision about how much to discount the welfare of future generations affects the present value of the benefits of tackling climate change today, and hence both the optimal pace of action and estimates of the cost of inaction. Although the quantitative conclusions change, the qualitative conclusions do not.

ACTIVITY 13.1

STERN VIEW OF DISCOUNT RATES

Figure 13.8 showed a 1000-year history of temperatures on the planet. Suppose we could all agree on the science of global warming. This would allow statements of the form, 'If we continue producing emissions at the current rate, global temperatures will rise according to the following profile, with the following consequences in terms of flooding, volatile weather, drought, and so on.'

Suppose too that there was only one country in the world, so we did not have to worry about whether the US or India participated in trying to slow down climate change. The central issue then would be, ‘how much pain should we inflict on today’s generation in order to mitigate the problem for future generations?’

The lower the discount rate we use in this calculation, the greater the present value of the benefits of helping future generations; the higher the discount rate we use, the less today we care about helping future generations. The Stern Review’s recommendation, that we should take urgent action to substantially reduce emissions, follows inexorably from its analysis provided we agree with its assumption that we should not discount the welfare of future generations in making this policy decision today.

Others, such as Professor William Nordhaus of Yale University, have argued that today’s decision makers should discount the welfare of future generations – not least because they are still likely to be richer than us and have better options than we face – in which case, the optimal policy response to climate change is a slower mitigation of emissions today, albeit then requiring that future generations will have to take much more drastic action.

The discount rate is not an academic abstraction. It affects key valuations and decisions, whether in the stock market or in the politics of controlling global warming.

Questions

- a. If we wish to weight equally the utility of current and future generations, what discount rate should we apply to future utility?
- b. Still weighting utility equally, suppose future generations are richer than us and we believe in the principle of diminishing marginal utility of consumption. Will a unit of consumption be worth more today when we are poor or more tomorrow when we are rich?
- c. Suppose, by sacrificing consumption today, we invest in physical capital that would make future generations richer. Say, on average, this investment has a rate of return of 5 per cent a year in real terms. What return would an environmental investment (for example, preventing climate change) have to yield in order for future generations to be pleased with the decisions we made today?

To check your answers to these questions, go to page 678.

13.7 Other missing markets: time and risk

The previous two sections were devoted to a single idea. When externalities exist, free market equilibrium is inefficient because the externality itself does not have a

market or a price. People take no account of the costs and benefits their actions inflict on others. Without a market for externalities, the price system cannot bring marginal costs and marginal benefits of these externalities into line. We now discuss other ‘missing markets’ – those for time and for risk.

The present and the future are linked. People save, or refrain from consumption, today in order to consume more tomorrow. Firms invest, reducing current output by devoting resources to training or building, in order to produce more tomorrow. How should society make plans today for the quantities of goods produced and consumed in the future? Ideally, everyone makes plans such that the social marginal cost of goods in the future just equals their social marginal benefit.

Chapter 12 discussed a *forward market*, in which buyers and sellers make contracts today for goods delivered in the future at a price agreed today. Suppose there is a forward market for copper in 2020. Consumers equate the marginal benefit of copper in 2020 to the forward price, which producers equate to the marginal cost of producing copper for 2020. With a complete set of forward markets for all commodities for all future dates, producers and consumers today make consistent plans for future production and consumption of all goods, and the social marginal benefit of every future good equals its social marginal cost.

Chapter 12 explained why few forward markets exist. You can trade gold but not cars or washing machines. Since nobody knows the characteristics of next year’s model of car or washing machine, we cannot write legally binding contracts to be easily enforced when the goods are delivered. Without these forward markets, the price system cannot equate the marginal cost and marginal benefits of planned future goods.

There are also few *contingent* or insurance markets for dealing with risk. People usually dislike risk. It reduces their utility. Does society undertake the efficient amount of risky activities?

A complete set of insurance markets lets risk be transferred from those who dislike risk to those who will bear risk at a price. The equilibrium price equates social marginal costs and benefits of risky activities. However, adverse selection and moral hazard inhibit the organization of private insurance markets. If some risky activities are uninsurable at any price, the price system cannot guide society to equate social marginal costs and benefits.

Future goods and risky goods are examples of commodities with missing markets. Like externalities, these are market failures. Free market equilibrium is generally efficient. And the theory of the second-best tells us that, when some markets are distorted, we probably do not want other markets to be completely distortion free.

13.8 Quality, health and safety

Information is incomplete because gathering information is costly. This leads to inefficiency. Consider a firm using benzene in its production process. A worker

unaware that exposure to benzene may cause cancer may work for a lower wage than if this information is widely available. The firm's production cost understates the true social cost and the good is overproduced. Governments regulate health, safety and quality standards because they recognize the danger of market failure.

UK examples include the Health and Safety at Work Acts, legislation to control food and drugs production, the Fair Trading Act governing consumer protection, and various traffic and motoring regulations. Such legislation aims to encourage the provision of information that lets individuals more accurately judge costs and benefits, and aims to set and enforce standards designed to reduce the risk of injury or death.

Providing information

Figure 13.11 shows the supply curve SS for a drug that is potentially harmful. DD is the demand curve if consumers do not know the danger. In equilibrium at E , the quantity Q is produced and consumed. With full information about the dangers, people would buy less of the drug. The demand curve DD' shows the marginal consumer benefit with full information. The new equilibrium at E' avoids the deadweight burden $E'EF$ from overproduction of the drug.

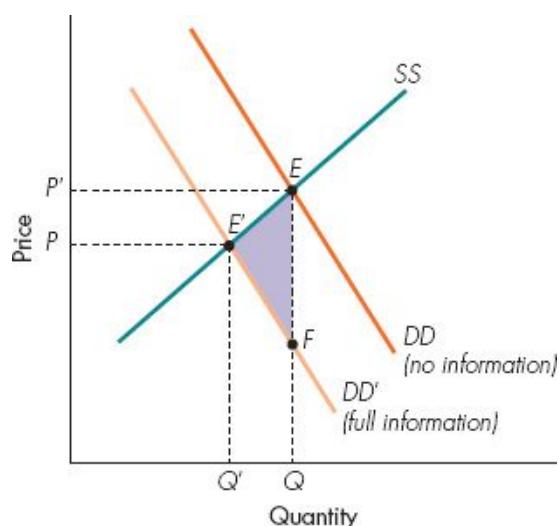


Figure 13.11 Information and unsafe goods

Consumers cannot individually discover the safety risks associated with a particular good. Free market equilibrium occurs at E . A government agency now provides information about the product. As a result, the demand curve shifts down and the new equilibrium is at E' , where the true or full information value of an extra unit of the good equals its marginal social cost. Providing information prevents a welfare cost $E'EF$ that arises when uninformed consumers use the wrong marginal valuation of the benefits of the good.

If information were free to collect, everyone would know the true risks. From the social gain $E'EF$, we should subtract the resources needed to discover this

information. Free market equilibrium is at E because it is not worth each individual checking up privately on each drug on the market. It makes sense for society to have a single regulatory body to check drugs, and a law whose enforcement entitles individuals to assume that drugs have been checked out as safe. Certification of safety or quality need not be carried out by the government. Sotheby's certify Rembrandts and the AA will check out a used car for you.

Two factors inhibit the use of private certification in many areas of health and safety. First, the public perceives a conflict between the profit motive and the incentive to tell the truth. Public officials may be less easily swayed. Second, a private certification agency might have to decide standards. What margin of error should be built into safety regulations? How safe must a drug be to get a certificate? These are issues of public policy. They involve externalities and have distributional implications. Even if society uses private agencies to *monitor* regulations, it usually sets the standards itself.

Imposing standards

The public interest is important when little is known about a product and where the consequences of any error may be catastrophic. Few believe that safety standards for nuclear power stations can be adequately determined by the private sector.

In imposing standards, governments raise the private cost of production by preventing firms from adopting the cost-minimizing techniques they otherwise would use. Sometimes the government has better information than the private sector. Sometimes standards compensate for externalities neglected by the private firm. Sometimes standards reflect a pure value judgement based on distributional considerations. One contentious area is the value of human life itself.

Politicians often claim that human life is beyond economic calculation and must be given absolute priority at any cost. An economist will make two points in reply. First, it is *impossible* to implement such an objective. It is too costly in resources to try to eliminate *all* risks of premature death. Sensibly, we do not go this far. Second, in occupational and recreational choices, for example driving racing cars or going climbing, people take risks. Society must ask how much more risk-averse it should be than the people it is trying to protect.

Beyond some point, the marginal social cost of further risk reduction exceeds the marginal social benefit. It takes a huge effort to make the world just a little safer, and the resources might have been used elsewhere to greater effect. Zero risk does not make economic sense. We need to know the costs of making the world a little safer, and we need to encourage society to decide how much it values the benefits. Not accounting for those costs and benefits may lead societies to choose inefficient allocations.

Summary

- **Welfare economics** deals with normative issues or value judgements. Its purpose is not to describe how the economy works but to assess how well it works.
- **Horizontal equity** is the equal treatment of equals, and **vertical equity** the unequal treatment of unequals. Equity is concerned with the distribution of welfare across people. The desirable degree of equity is a pure value judgement.
- A **resource allocation** is a complete description of what, how and for whom goods are produced. To separate as far as possible the concepts of equity and efficiency, economists use Pareto efficiency. An allocation is **Pareto-efficient** if no reallocation of resources would make some people better off without making others worse off. If an allocation is inefficient it is possible to achieve a Pareto gain, making some people better off and none worse off. Many reallocations make some people better off and others worse off. We cannot say whether such changes are good or bad without making value judgements to compare different people's welfare.
- For a given level of resources and a given technology, the economy has an infinite number of Pareto-efficient allocations that differ in the distribution of welfare across people. For example, every allocation that gives all output to one individual is Pareto-efficient. But there are many more allocations that are inefficient.
- Under strict conditions, competitive equilibrium is Pareto-efficient. Different initial distributions of human and physical capital across people generate different competitive equilibria corresponding to each possible Pareto-efficient allocation. When price-taking producers and consumers face the same prices, marginal costs and marginal benefits are equated to prices (by the individual actions of producers and consumers).
- Governments face a conflict between equity and efficiency. Redistributive taxation drives a wedge between prices paid by consumers (to which marginal benefits are equated) and prices received by producers (to which marginal costs are equated). Free market equilibrium will not equate marginal cost and marginal benefit and there will be inefficiency.
- **Distortions** occur whenever free market equilibrium does not equate **marginal social cost** and **marginal social benefit**. Distortions lead to inefficiency or **market failure**. Apart from taxes, there are three other important sources of distortion: imperfect competition (failure to set price equal to marginal cost), externalities (divergence between private and social costs or benefits), and other

missing markets in connection with future goods, risky goods or other informational problems.

- When only one market is distorted the **first-best** solution is to remove the distortion, thus achieving full efficiency. The first-best criterion relates only to efficiency. Governments caring sufficiently about redistribution might still prefer inefficient allocations with more vertical equity. However, when a distortion cannot be removed from one market, it is not generally efficient to ensure that all other markets are distortion-free. The theory of the **second-best** says that it is more efficient to spread inevitable distortions thinly over many markets than to concentrate their effects in a few markets.
- **Production externalities** occur when actions by one producer directly affect the production costs of another producer, as when one firm pollutes another's water supply. **Consumption externalities** mean one person's decisions affect another consumer's utility directly, as when a garden gives pleasure to neighbours.
- Externalities lead to divergence between private and social costs or benefits because there is no implicit market for the externality itself. When only a few people are involved, a system of **property rights** may establish the missing market. The direction of compensation will depend on who has the property rights. Either way, it achieves the efficient quantity of the externality at which marginal cost and marginal benefit are equated. The efficient solution is rarely a zero quantity of the externality. **Transaction costs** and the **free-rider problem** may prevent implicit markets being established. Equilibrium will then be inefficient.
- When externalities lead to market failure, the government could set up the missing market by pricing the externality through taxes or subsidies. If it were straightforward to assess the efficient quantity of the externality and hence the correct tax or subsidy, such taxes or subsidies would allow the market to achieve an efficient resource allocation.
- In practice, governments often regulate externalities such as **pollution** or **congestion** by imposing standards that affect quantities directly rather than by using the tax system to affect production and consumption indirectly. Overall quantity standards may fail to equate the marginal cost of pollution reduction across different polluters, in which case the allocation will not be efficient. However, simple standards may use up fewer resources in monitoring and enforcement and may prevent disastrous outcomes when there is uncertainty.

- **Global warming** represents a negative environmental externality that is believed to be caused by human activity and is posing a serious threat to the global economy. As an externality, it has two main dimensions: an intergenerational dimension and an international dimension. To assess the possible effects of global warming on our economies, we employ cost–benefit analysis.
- **Moral hazard, adverse selection and other informational problems** prevent the development of a complete set of **forward markets** and **contingent markets**. Without these markets, the price system cannot equate social marginal costs and benefits for future goods or risky activities.
- **Incomplete information** may lead to inefficient private choices. Health, quality and safety regulations are designed both to provide information and to express society’s value judgements about intangibles, such as life itself. By avoiding explicit consideration of social costs and benefits, government policy may be inconsistent in its implicit valuation of health or safety in different activities under regulation.

Review questions



EASY

- 1 An economy has 10 units of goods to share out between 2 people. $[x, y]$ means that the first person gets a quantity x , the second person a quantity y . For each of the allocations (a) to (e), say whether they are (i) efficient and (ii) equitable: (a) [10, 0], (b) [7, 2], (c) [5, 5], (d) [3, 6], (e) [0, 10]. What does ‘equitable’ mean? Would you prefer allocation (d) to allocation (e)?
- 2 John and Jennifer need to decide how to divide a chocolate cake of size one. Putting the quantity of the cake that John can get on the vertical axis and the quantity of the cake that Jennifer can get on the horizontal axis, plot the Pareto frontier of this cake allocation problem. What does a point below the Pareto frontier represent? Is an allocation where John gets the entire cake and Jennifer nothing Pareto-efficient?
- 3 In deciding to drive a car in the rush hour, you think about the cost of petrol and the time of the journey. Do you slow other people down by driving? Is this an externality? Will too many or too few people drive cars in the rush hour? Should commuter parking in cities be restricted?
- 4 **Common fallacies** Why are these statements wrong? (a) Society should ban all toxic discharges. (b) Anything governments can do, the market can do better. (c) Anything the market can do, the government can do better.
- 5 In 1885, 200 people died when the steam boiler exploded on a Mississippi river boat. Jeremiah Allen and three friends formed a private company offering to insure any boiler that they had inspected for safety. Boiler inspections caught on,

and explosion rates plummeted. Would Jeremiah Allen's company have been successful in reducing explosion rates if it had certified boilers but not insured them as well? Explain.

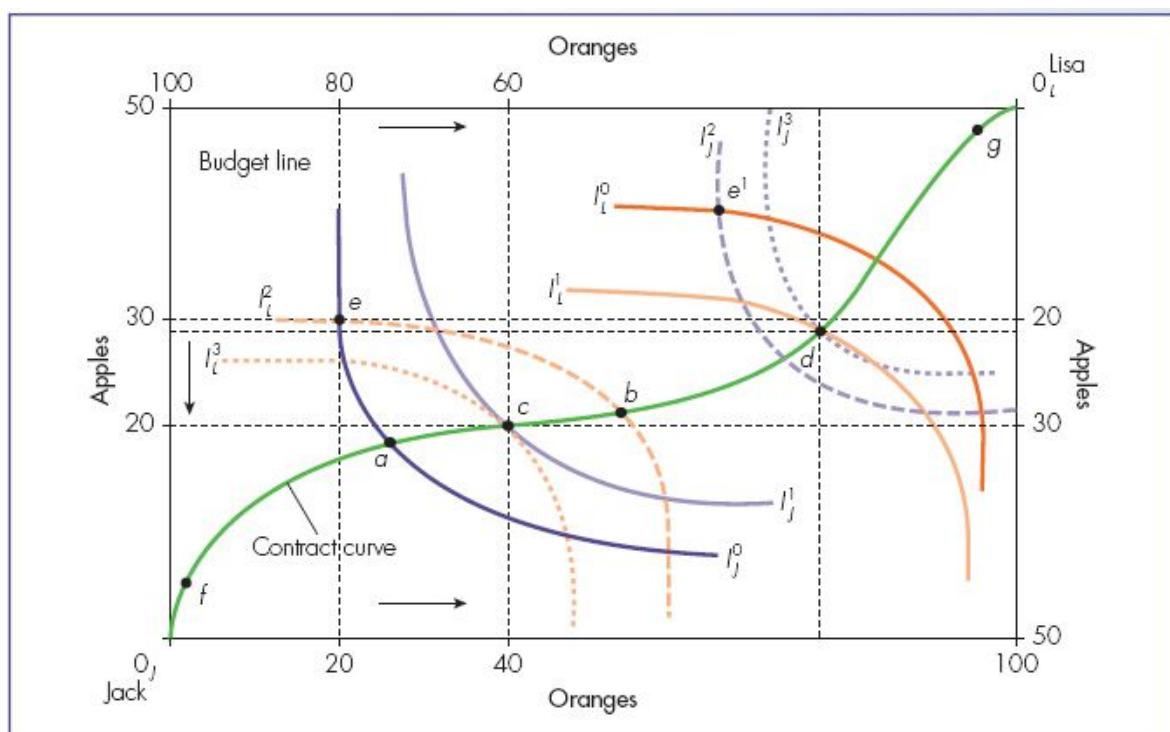
MEDIUM

- 6 The price of meals is £1 and the price of films is £5. There is perfect competition in the market and no externalities exist. Labour is the variable factor of production and workers can move freely between industries. What can we say about (a) the relative benefit to consumers of a marginal film and a marginal meal, (b) the relative marginal production costs of films and meals, and (c) the relative marginal product of variable factors in the film and meal industries? Why is this equilibrium efficient?
- 7 A government needs to raise £10 billion from taxes. It knows that taxes create deadweight losses, and it taxes a number of activities and products. For the most efficient outcome possible, should the tax rate on each of the following be low or high: (a) alcohol, (b) branded clothing, (c) caviar, (d) tobacco?
- 8 A honey firm is located next to an apple field owned by a farmer. The bees go into the apple field and help to make all the trees more productive. This in turn reduces the costs of the farmer. We have a positive externality. The total cost of the honey firm is $[TC_{\{H\}}] = [H]^2$, where H is the amount of honey. For the farmer the total cost is $[TC_{\{A\}}] = [A^2] - [H]$, where A is the amount of apples. Assume that the price of honey is fixed at £2, while the price of an apple is fixed at £4. Write down the profit functions for the honey firm and for the farmer. What is the profit-maximizing level of honey produced by the honey firm? What is the profit-maximizing level of apples produced? Find the profits earned by the honey firm and the farmer.
- 9 Now suppose that the honey firm and the farmer in Question 8 merge to become a single firm that produces honey and apples. The total cost faced by the merged firm is $[TC_{\{M\}}] = [H^2] + [A^2] - [H]$. The prices of the two goods are the same as in Question 8. Write down the profit function of the merged firm. Find the profit-maximizing levels of honey and apples produced. What is the total profit obtained by the merged firm? Compare your answer with your results in Question 8. Is the externality internalized?
- 10 Much of the economics of efficiency is about ensuring that we equate the marginal cost of producing the last unit with the marginal benefit of that unit to the last consumer. Suppose the marginal cost of preventing the planet from overheating is £10 000 billion. How would you attempt to assess the marginal benefit?
- 11 (a) Why might society ban drugs that neither help nor harm the diseases they are claimed to cure? (b) If regulatory bodies are blamed for bad things that happen despite regulations (a train crash) but not blamed for preventing good things through too much regulation (rapid availability of a safe and useful drug), will regulatory bodies over-regulate activities under their scrutiny?

HARD

|2 Suppose Jack has 30 apples and 20 oranges and Lisa has 20 apples and 80 oranges. So, there are 100 oranges and 50 apples in total. There is no production activity. Jack and Lisa must decide between consuming their initial endowment of the two goods or trading with each other. Given below is the Edgeworth box showing Jack's and Lisa's allocation of resources. Jack's indifference curves are I^0J , I^1J , I^2J , I^3J and Lisa's indifference curves are I^0L , I^1L , I^2L , I^3L .

- Identify the Pareto-efficient allocation of the two goods between Jack and Lisa.
- Now suppose the initial endowment is at point e^1 . Explain how government intervention can bring about changes in the allocation such that it is not only Pareto-efficient but also equitable.



|3 **Essay question** Why do politicians pretend that trains can be made perfectly safe and hospitals can supply all the health care that we know how to supply, when it is perfectly obvious that we do not have the resources to do these things and that it would be highly wasteful to try?

|4 A firm producing plastic bags is polluting the air in a neighbourhood. In the following table, the marginal private costs (MPC) of the firm for different quantities of plastic bags are reported, together with the price of plastic bags.

Q	$MPC (\text{£})$	Selling price
1	11	28
2	12	26
3	13	24
4	14	22
5	15	20

6	16	18
7	17	16
8	18	14
9	19	12

Polluting the air creates an externality. We know that the value of the externality is £10 for each quantity level. In a graph with Q on the horizontal axis, plot the marginal social cost (MPC) and the demand. Show the equilibrium in the market. Why is the equilibrium inefficient?

- 1 James J. Heckman shared the Nobel Prize in Economics in 2000 with Daniel McFadden.
- 2 Lump sum taxes are fixed amounts that do not change with changes in the entity taxed. They are an example of non-distortionary taxes.
- 3 The results of the analysis will not change if we consider a positively sloped MPC curve. What matters is that marginal private costs lie below the marginal social costs.
- 4 Conversely, a farmer who spends money on pest control reduces pests on nearby farms. If production externalities are beneficial, the marginal social cost is below the marginal private cost. Suppose we swap the labels MSC and MPC in Figure 13.5. Free market equilibrium is at E but E is now the efficient allocation.
- 5 From Ronald Harry Coase, Nobel Prize winner in Economics in 1991, who first pointed out the relationship between property rights and efficiency in the presence of externalities.
- 6 The free-rider problem is an important issue in the provision of public goods. This will be discussed in Chapter 14.

CHAPTER 14

Government spending and revenue

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 define different kinds of government spending
- 2 understand why public goods cannot be provided by a market
- 3 identify average and marginal tax rates
- 4 understand how taxes can compensate for externalities
- 5 define supply-side economics
- 6 understand why tax revenue cannot be raised without limit
- 7 recognise how cross-border flows limit national economic sovereignty
- 8 understand the political economy within which governments set policy

The scale of government rose steadily until the 1970s. Then many people felt it had become too big, using resources better employed in the private sector. High taxes were thought to be stifling private enterprise. Electorates in many countries turned to the political leaders who promised to reduce the scale of government.

Now the pendulum is swinging back. After the financial crisis of 2007–08, many developed economies faced a recession. The electorate sees a reduction in public spending as counterproductive, even in countries with a high burden of public debt.

For historical perspective, Table 14.1 shows how government grew everywhere in the last century.

Table 14.1 Government spending (% of GDP)

	1880	1960	2012
Japan	11	18	33.8
US	8	28	35.0
Germany	10	32	45.0
UK	10	32	48.5
France	15	35	56.6
Sweden	6	31	52.0

Sources: EUROSTAT; www.bea.gov; *CIA World Factbook*.

Most government spending is financed by tax revenue. However, just as you may overspend your student income by borrowing now and repaying later, the government need not balance its spending and revenue in any particular period. When the difference between total revenues and total spending is negative, we then have a budget deficit. When that difference is positive, the government is running a **budget surplus**. Table 14.2 shows that, by 2012, the US and the UK had budget deficits higher than in France while Germany had a budget surplus.

After this broad background, we now examine microeconomic issues. First, we distinguish between **marginal and average tax rates**.

In a *progressive* tax structure, the average tax rate rises with an individual's income. The government takes proportionately more from the rich than from the poor. In a *regressive* tax structure, the average tax rate falls as income level rises, taking proportionately less from the rich.

Table 14.3 shows that the UK, like most countries, has a progressive income tax structure. Figure 14.1 explains why. We plot pre-tax income on the horizontal axis and post-tax income on the vertical axis. The line OG, with a slope of 45 degrees, implies no taxes. A pre-tax income *OA* on the horizontal axis matches the same post-tax income *OA* on the vertical axis. Now suppose there is an income tax, but the first *OA* of income is untaxed. If the marginal tax rate on taxable income is constant, individuals face a schedule *OBCD*, keeping a constant fraction of each pound of pre-tax income above OA. The higher the marginal tax rate, the flatter is BC.

The **budget surplus** (deficit) is the excess (shortfall) of government's spending over its revenue.

The **marginal tax rate** is the fraction of the last pound of income paid in tax.

The **average tax rate** is the fraction of total income paid in tax.

Table 14.2 Government activity in 2012 (% of GDP)

	UK	US	France	Germany
Spending	48.5	35.0	56.6	45.0
Total revenue	42.2	27.5	51.7	45.2
Budget deficit	26.3	27.5	24.9	0.2

Sources: EUROSTAT; www.bea.gov.

To calculate the average tax rate at a point such as D, we join up OD. The flatter the slope of this line, the higher is the average tax rate. Even with a constant marginal tax rate, and a constant slope along BC, the initial tax allowance makes the tax structure progressive. The higher an individual's gross income, the smaller is the tax allowance as a percentage of gross income, so the larger is the fraction of total income on which tax is paid.

But Table 14.3 shows that *marginal* tax rates may also rise with income. As individuals move into higher tax bands they pay higher marginal tax rates, moving on to even flatter portions of the tax schedule. The average tax rate now rises sharply with income.

Table 14.3 shows that UK marginal tax rates have fallen a lot in the past two decades, especially for the very rich. A millionaire paying an 83 per cent tax rate on all taxable income except the first £70 000 in 1978 paid only 40 per cent in 2008/09.

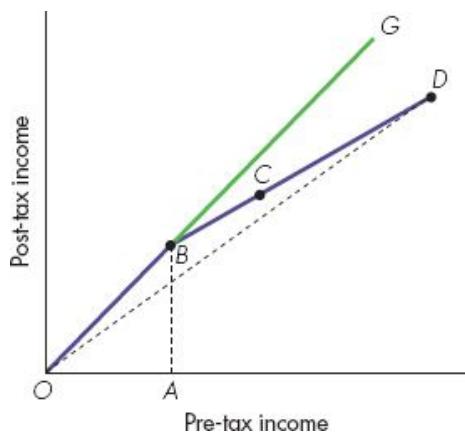
Table 14.3 UK income tax rates, 1978–2008

Taxable income (2004 £000s)	Marginal tax rate (%)	
	1978/79	2008/09
2020	34	20
5000	34	20
10 000	34	20
20 000	45	20
31 400	50	20
40 000	70	40
70 000	83	40

Note: Taxable income after deducting allowances. In 2008/09 a single person's allowance was almost £5500.

Sources: HMSO, *Financial Statement and Budget Report*; ONS, *Budget 2007*.

The UK was not alone in cutting tax rates. There was a worldwide move to cut tax rates, especially for the very rich. In part, this reflected the belief that tax rates were previously so high that distortions had been large.



The 45° line OG shows zero taxes or transfers so that pre-tax and post-tax income coincide. With an allowance OA , then a constant marginal tax rate t , the post-tax income schedule is $OBCD$. The slope depends only on the marginal tax rate [on BCD it is $(1 - t)$]. The average tax rate at any point D is the slope of OD . A tax is progressive if the average tax rate rises with pre-tax income.

Figure 14.1 A progressive income tax

However, it also reflected increasing competition between governments to attract mobile resources (physical and human capital) to their country. At the end of the chapter we discuss how cross-border mobility undermines national sovereignty.

14.1 Taxation and government spending

Government spending, and the taxes that finance it, are now about 45 per cent of national output. Figure 14.2 shows the composition of government spending and revenue in 2011/12. Total expenditure in 2011/12 was around £710 billion, while revenues were around £589 billion.

A **transfer payment** requires no good or service in return during the period in which it is paid.

Direct taxes are taxes on income and wealth.

Indirect taxes are taxes on spending and output.

More than a third of total government spending went on **transfer payments** such as social protection of pensions, jobseeker's allowance (formerly unemployment benefit) and debt interest. Of the remaining spending directly on goods and services, the most important spending categories are health, defence and education. Figure 14.2 also shows how this government spending is financed. The most important **direct taxes** are income tax, and corporation tax on company profits.

The most important **indirect taxes** are value added tax (VAT) and customs duties. Note that, since state provision of retirement pensions is included on the expenditure side as a transfer payment, pension contributions under the national insurance scheme are included on the revenue side.

A **public good** is a good for which individuals cannot be excluded from using it and the use by an individual does not reduce availability to others.

A **private good**, if consumed by one person, cannot be consumed by others.

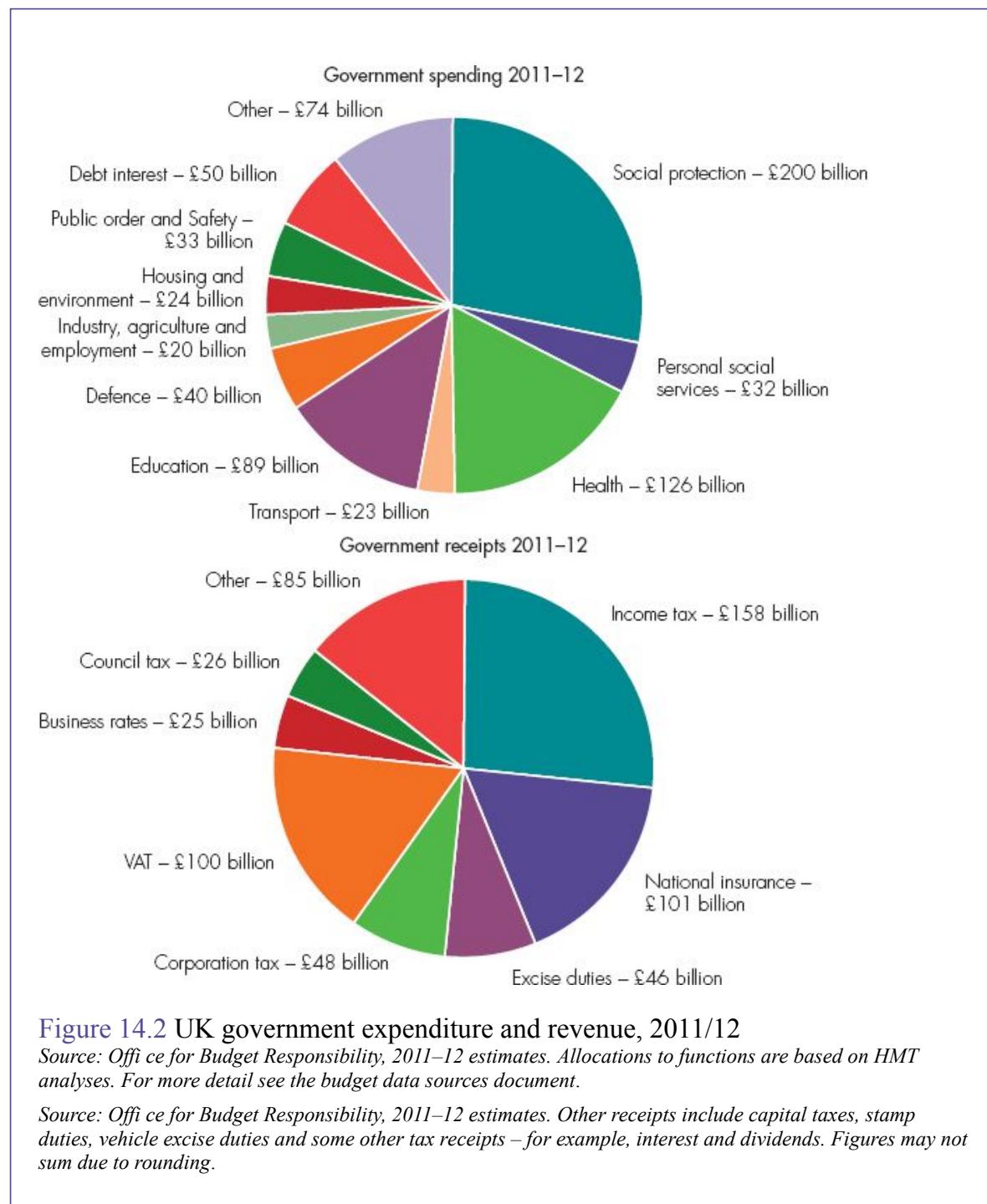
14.2 The government in the market economy

How do we justify government spending in a market economy?

Public goods

In Chapter 13 we introduced the idea of **public good** as market failure. Here, we discuss this issue in more detail.

Ice cream is a **private good**. If you eat an ice cream nobody else can eat that particular ice cream. For any given supply, your consumption reduces the quantity available for others to consume. Most goods are private goods.



Clean air and are examples of public goods. If the air is pollution-free, your consumption of it does not interfere with our consumption of it. If the navy is patrolling coastal waters, your consumption of national defence does not affect our quantity of national defence. We all consume the same quantity; namely, the quantity is supplied in the aggregate. We may get different amounts of utility if our tastes differ, but we all consume the same quantity.

The key aspects of public goods are:

1. They are *non-rivalrous*: it is technically possible for one person to consume without reducing the amount available for others.
2. They are *non-excludable*: it is impossible to exclude anyone from consumption except at a prohibitive cost.

A football match can be watched by many people, especially if it is on TV, without reducing the quantity consumed by other viewers; but *exclusion* is possible. The ground holds only so many, and some Premier League clubs now charge to watch their games live on their own TV stations. The interesting issues arise when, as with national defence, exclusion of certain individuals from consumption is impossible.

Free-riders

Chapter 13 introduced the *free-rider problem* when discussing why bribes and compensation for externalities might not occur. Public goods are wide open to the free-rider problem if they are supplied by the private sector. Since you get the same quantity of national defence as everyone else, *whether or not you pay for it*, you never buy national defence in a private market. Nor does anyone else. No defence is demanded, even though we all want it.

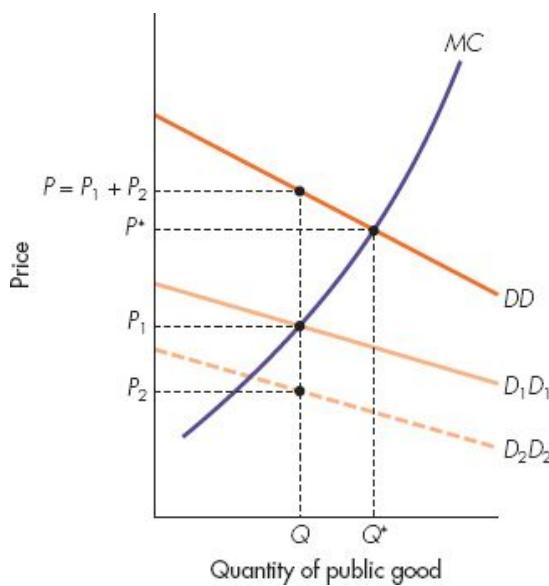
Public goods are like a strong externality. If you buy defence, everyone else also gets the benefits. Since marginal private and social benefits diverge, private markets will not produce the socially efficient quantity. Government intervention is needed.

The marginal social benefit

Suppose the public good is a pure public water supply. The more infected the water, the more people are likely to get cholera. Figure 14.3 supposes there are two people. The first person's demand curve for water purity is D_1D_1 . Each point on the demand curve shows what she would pay for the last unit of purer water; that is, her marginal benefit. D_2D_2 shows the marginal benefit of purer water to the second person.

Curve DD is the marginal social benefit of purer water. At each level of the public good, we *vertically* sum the marginal benefit of each individual to get the marginal social benefit. At the output Q , the marginal social benefit is $P = P_1 + P_2$. We sum vertically at a given quantity because everyone consumes the same quantity of a public good.

Figure 14.3 also shows the marginal cost of the public good. If there are no production externalities, the marginal private cost and marginal social cost coincide. The socially efficient output of the public good is Q^* , where the marginal social benefit equals the marginal social cost.



D_1D_1 and D_2D_2 are the separate demand curves of two individuals and show the marginal private benefit of the last unit of the public good to each individual. What is the social marginal benefit of the last unit to the group as a whole? Since both individuals consume whatever quantity of the good is produced, we must add up vertically the price each is prepared to pay for the last unit. At the output Q the marginal social benefit is thus $P_1 + P_2$. The curve DD showed the marginal social benefit and is obtained by vertically adding the demand curves of the two individuals. If MC is the private and social marginal cost of producing the public good, the socially efficient output is at which social marginal cost and social marginal benefit are equal.

Figure 14.3 A pure public good

What happens if the good is privately produced and marketed? Person 1 might pay P_1 to have a quantity Q produced by a competitive supplier pricing at marginal cost. At the output Q , the price P_1 just equals the marginal private benefit that person 1 gets from the last unit of the public good. Person 2 will not pay to have the output of the public good increased beyond Q . Person 2 cannot be excluded from consuming the output Q that person 1 has commissioned. At the output Q , person 2's marginal private benefit is only P_2 , less than the current price P_1 . Person 2 will not pay the higher price needed to induce a competitive supplier to expand output beyond Q . Person 2 free-rides on person 1's purchase of Q . This quantity privately produced and consumed in a competitive market is below the efficient quantity Q^* .

Revelation of preferences

If it knows the marginal social benefit curve DD , the government can decide the efficient output of the public good. How does the government discover the individual demand curves that must be vertically added to get DD ? If people's payments for the good are related to their individual demand curves, everyone will lie. People will understate how much they value the good in order to reduce their own payments, just as in a private market. Conversely, we are all for safer streets if we do not have to contribute to the cost.

In practice, democracies try to resolve this problem through election of governments. Politics lets society get closer to the efficient answer than the market can. Different parties offer different quantities of public goods, together with statements on how they will be financed by taxes. By asking ‘How much would you like, given that everyone is charged for the cost of providing public goods?’ society comes closer to providing the efficient quantities of public goods. However, with only a few parties competing in an election and many policies on which they offer a position, this remains a crude way to decide the quantities of public goods provided.

Government production

The output of public goods must be *decided* by the government, not the market. This need not mean government must produce the goods itself. Public goods need not be produced by the public sector.

National defence is a public good largely produced in the public or government sector. We have few private armies. Street-sweeping, though a public good, can be subcontracted to private producers, even if local government determines its quantity and pays for it out of local tax revenue. Conversely, state hospitals involve public sector production of private goods. One person’s hip replacement operation prevents the busy surgeon from operating on someone else.

In the next chapter we examine why the public sector may wish to produce private goods. Whether public goods need to be produced by the public sector depends not on their consumption characteristics, on which our definition of public good relies, but on their production characteristics. There is nothing special about street-sweeping. In contrast, armies rely on discipline and secrecy. Generals and admirals may believe, and society may agree, that offences against these regulations should receive unusual penalties not generally sanctioned in private firms. Few people believe that insubordination is an important offence for street-sweepers.

CASE 14.1

THE PARADOX OF OPEN SOURCE SOFTWARE

Open source software, developed by volunteers, represents a case of a public good that is somehow paradoxical. It is a public good since the ‘source code’ used to generate the programs is freely available -hence ‘open source’.

According to the theory of public goods, without government intervention, a public good will not be provided at the efficient level. However, open source software is now quite popular. Why have private agents, without property rights over the source code, invested effort in developing a public good subject to free-riding?

A possible answer may be that such agents are moved by reputation building and career concerns. A software programmer who is able to prove his skills by programming open source code may have a chance to be employed by important software companies. Therefore, according to this view, private agents have an

incentive to provide effort in developing open source software since this will signal their quality as a programmer.

Another possible answer is that open source software is not an alternative to proprietary software (like Microsoft, which has the property rights for its software), but instead can be viewed as a complement.

Proprietary provision fails to effectively meet the needs of many customers in markets where customers have highly disparate needs and products are complex. Open source software and proprietary provision of pre-packaged software can both exist in a market, recognizing that they mainly serve different groups of customers. Open source will be used most by firms that have their own development capability and which have complex, specialized needs; pre-packaged software will be used by firms with simpler needs and those which lack development capabilities.

Source: J. Bessen, 'Open source software: Free provision of complex public goods', working paper, Boston University School of Law and Research on Innovation, 2005.

Transfer payments and income redistribution

Government spending on transfer payments is primarily concerned with *equity* and *income redistribution*. By spending money on the unemployed, the old and the poor, the government alters the distribution of income and welfare that a free market economy would otherwise have produced: there is a minimum standard of living below which no citizen should fall. The specification of this standard is a pure value judgement.

To finance this spending, the government taxes those who can afford to pay. Taken as a whole, the tax and transfer system takes money from the rich and gives it to the poor. The poor get cash transfers but also enjoy the consumption of public goods paid for by income taxes raised from the rich. Figure 14.4 shows the effects of government intervention during 2009/10. The net taxes paid by the richest 10 per cent of the population amount for almost half of their disposable income. In contrast, the poorest 10 per cent of the population benefited from changes to the tax and benefit by an amount almost equal to half of their disposable income.

The desirable amount of redistribution is a value judgement on which people and parties will disagree. There is also the trade-off between efficiency and equity. To redistribute more the government has to raise tax rates, driving a larger wedge between the price paid by the purchaser and the price received by the seller. Since the price system achieves efficiency by inducing each individual to equate marginal cost or marginal benefit to the price received or paid, and hence to one another, taxes are generally distortionary and reduce efficiency.

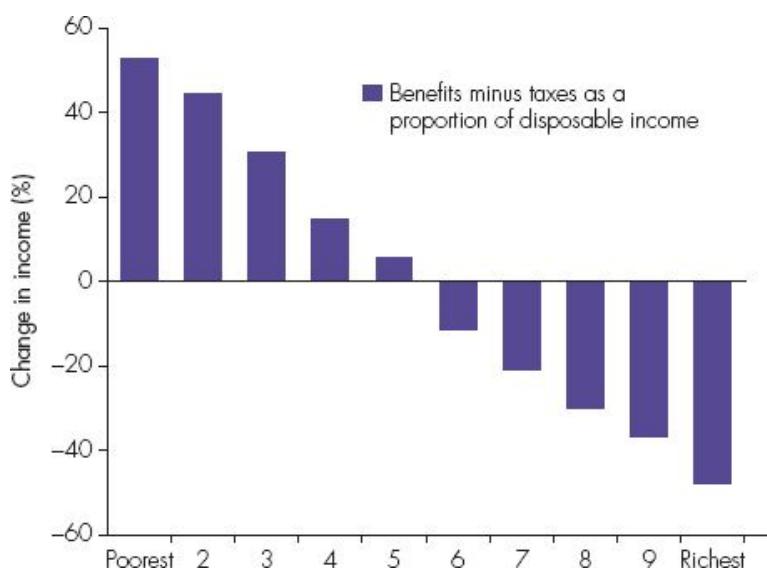


Figure 14.4 Distributional impact of the UK tax and benefits in 2009-2010 (% of initial disposable income)

Source: Redistribution, Work Incentives and Thirty Years of UK Tax and Benefit Reform working paper by Stuart Adam and James Browne, © Institute for Fiscal Studies, 2010, <http://www.ifs.org.uk/wps/wp1024.pdf>

Merit and demerit goods

Merit goods are goods that societies should consume regardless of whether an individual wants them. Those goods are provided by governments despite the fact that they can be consumed and bought individually in the market. The reason is that, if we leave to the private market the burden of providing merit goods, those goods will be underprovided. Merit goods include education and health.

On the other hand, there are also **demerit goods**. Demerit goods include cigarettes and heroin. Since society places a different value on these goods from the value placed on them by the individual, individual choice in a free market leads to a different allocation from the one that society wishes.

Merit (demerit) goods are goods that society thinks everyone should have (not have) regardless of whether an individual wants them.

There are two reasons for providing merit goods. The first is externalities. Indeed, merit goods generate positive externalities. If more education raises the productivity not merely of an individual worker but also of the workers with whom he co-operates, he ignores this production externality when choosing how much education to acquire. If people demand too little education, society should encourage the provision of education.

Conversely, if people ignore the burden on state hospitals when deciding to smoke and damage their health, society may regard smoking as a demerit to be discouraged. Taxing cigarettes may offset externalities that individuals fail to take into account.

The second reason for providing merit goods is that a society may believe that individuals no longer act in their own best interests and so it has to decide on their behalf. Addiction to drugs, tobacco or gambling are examples. Economists rarely subscribe to paternalism. The function of government intervention is less to tell people what they ought to like than to allow them better to gain what they already like. However, the government sometimes has more information or is in a better position to take a decision. Many people hate going to school, but later are glad they did. The government may spend money on compulsory education or compulsory vaccination because it recognizes that otherwise individuals act in a way they will subsequently regret.

14.3 The principles of taxation

This section is in three parts. First, we consider different taxes through which the government can raise revenue. Then we consider equity implications of taxation. Finally, we examine efficiency implications of taxation.

Types of taxes

Governments can collect tax revenue only if they monitor and enforce the activities being taxed. Before sophisticated records of income or sales, governments raised most of their revenue from customs duties and road tolls, places where transactions were easily monitored. Income tax in peacetime was not introduced in the UK until the 1840s, and VAT not until the 1970s.

How to tax fairly

The last chapter gave two notions of equity: *horizontal equity*, or the equal treatment of equals, and *vertical equity*, the redistribution from the ‘haves’ to the ‘have-nots’.

Progressive taxes reflect the principle of *ability to pay*. The principle of ability to pay reflects a concern about vertical equity. Thus, car users should be taxed to finance public roads. However, the **benefits principle** often conflicts with the principle of ability to pay. If those most vulnerable to unemployment pay the highest contributions to a government unemployment insurance scheme, it is hard to redistribute income or welfare. If the main objective is vertical equity, ability to pay must take precedence.

The **benefits principle** is that people getting most benefit from public spending should pay most for it.

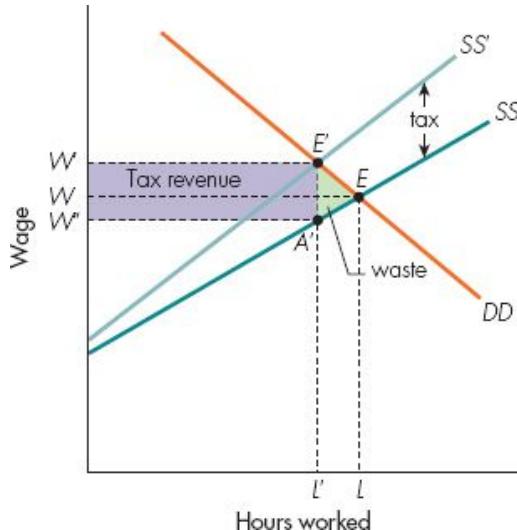
Two factors make the entire tax and benefit structure more progressive than an examination of income tax alone would suggest. First, transfer payments actually give money to the poor. The old receive pensions, the unemployed receive unemployment benefit and, as a final safety net, anyone whose income falls below a certain minimum is entitled to supplementary benefit. Second, the state provides public goods that can be consumed by the poor, even if they have not paid any taxes to finance these goods.

However, the system of tax, transfer and spending has some *regressive* elements that take proportionately more from the poor. Beer and tobacco taxes are huge earners for the government. Yet the poor spend a much higher proportion of their income on these goods than do the rich. Regressive taxes inhibit redistribution from the rich to the poor.

Tax incidence is the final tax burden once we allow for all induced effects of a tax.

Tax incidence

The ultimate effect of a tax can be very different from its initial effect. Figure 14.5 shows the market for labour. DD is the demand curve and SS the supply curve. Without an income tax (a tax on wages), labour market equilibrium is at E .



With no tax, equilibrium is at E and the wage is W . A wage tax raises the gross wage paid by firms above the net wage received by workers. Measuring gross wages on the vertical axis, the demand curve DD is unaltered by the imposition of the tax. Firms demand labour to equate the gross wage to the marginal value product of labour. SS continues to show labour supply, but as a function of the net wage. To get labour supply in terms of the gross wage we draw the new supply curve SS' . SS' lies vertically above SS by a distance reflecting the tax on earnings from the last hour worked. The new equilibrium is at E' . The hourly wage paid by firms is W' but the net wage received by workers is W'' . The vertical distance $A'E'$ shows the tax rate. Whether the government collects the tax from firms or from workers, the incidence of the tax is the same. It falls partly on firms, who pay a higher gross wage W' and partly on workers, who receive the lower net wage W'' . The area of pure waste $A'E'E$ is discussed in the text.

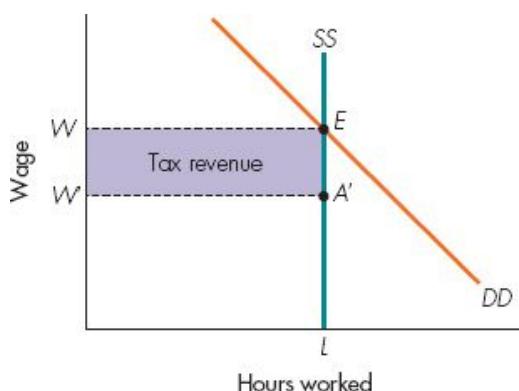
Figure 14.5 A tax on wages

Now the government imposes an income tax. If we measure the gross wage on the vertical axis, the demand curve DD is unaltered. Firms' demand for labour depends on the gross wage that they pay. Workers' preferences are unchanged, but it is the wage net-of-tax that workers compare with the marginal value of their leisure in deciding how much labour to supply. SS continues to show labour supply in terms of the net-of-tax wage, but we must draw in the higher schedule SS' to show the supply of labour in terms of the gross or *tax-inclusive* wage. The vertical distance between SS' and SS is the tax on earnings from the last hour's work.

DD and SS' show the behaviour of firms and workers at any gross wage. At the new equilibrium E , the gross wage is W' and firms demand L' workers. The vertical distance between A' and E is the tax paid on the last hour of work. The net-of-tax wage is W^2 , at which workers supply L' hours.

The tax on wages has raised the pre-tax wage to W' , but lowered the after-tax wage to W^2 . It has raised the wage that firms pay but lowered the take-home wage for workers. The incidence of the tax falls on *both* firms and workers.

Figure 14.6 shows the extreme case in which supply is completely inelastic. With no tax, equilibrium is at E and the wage is W . Since the vertical supply curve SS means that a fixed quantity of hours L is supplied whatever the after-tax wage, a tax on wages leads to a new equilibrium at A' . Only if the gross wage is unchanged will firms demand the quantity L that is supplied. Hence the entire incidence falls on the workers.



If the supply curve SS is vertical, a tax $A'E$ per unit leaves the quantity L unaffected. Since the demand curve DD is unaltered, the tax has no effect on the pre-tax wage rate. The full incidence of the tax falls on workers whose after-tax wage is reduced by the full amount of the tax.

Figure 14.6 Taxing a factor in inelastic supply

We can draw one very general conclusion. The more inelastic the supply curve and the more elastic the demand curve, the more the final incidence will fall on the seller rather than on the purchaser.

To check you have grasped the idea of incidence, draw for yourself a market with an elastic supply curve and an inelastic demand curve. Show that the incidence of a tax will now fall mainly on the purchaser.¹

Taxation, efficiency and waste

Taxes have efficiency effects as well as equity effects. We can use Figure 14.5 again. Before the tax is imposed, labour market equilibrium is at E . The wage W measures both the marginal social benefit of the last hour of work and its marginal social cost. The demand curve DD tells us the marginal benefit of the extra goods produced. The supply curve SS tells us the marginal value of the leisure being sacrificed in order to work another hour; that is, the marginal social cost of extra work. At E , marginal social costs and benefits are equal, which is socially efficient.

When the tax is imposed, the new equilibrium is at E' ¹. The tax $A'E'$ increases the wage to firms to W' but reduces the after-tax wage for workers to W^2 . But there is an additional tax burden or deadweight loss that is pure waste. It is the triangle AEE . By reducing the quantity of hours from L to L' , the tax drives a wedge between marginal benefit, the height of the demand curve DD , and marginal social cost, the height of the supply curve SS . This distortion makes free market equilibrium inefficient.

CASE 14.2

DO YOU MIND IF I SMOKE? THE SMOKING BAN IN THE UK

The smoking ban that took effect in the UK in 2006/07 is an example of a government policy to tackle a negative externality. The ban makes it illegal to smoke in all enclosed public venues and workplaces in the UK.

Smoking is a negative externality since smokers pollute the air for other people but ignore this in deciding how much to smoke. Doing so has negative effects on passive smokers. Doctors estimate that second-hand smoke kills more than 600 people a year.

Moreover, it has negative effects on society as a whole. By smoking, smokers have a greater likelihood of suffering from smoking-related diseases and so they are likely to need health care in the future. This will affect health care expenditure in the country. Since the public health care system is financed also by non-smokers, smoking will have a negative effect on non-smokers as well.

Finally, smoking is viewed as a demerit good and therefore governments should do something to discourage people from smoking. The smoking ban, together with heavy taxation on cigarettes, aims to reduce the number of smokers so that the negative externality created by smoking will decrease.

After the introduction of the smoking ban, cigarette sales decreased by 7 per cent. By looking at markets linked to the cigarette market, we can also gain a better idea of the general equilibrium effect of the smoking ban. In particular, we can look at the sales figures for public houses. Market research organization Nielsen estimated that beer sales in England and Wales could drop by 200 million pints each year as a result of the ban.

Source: adapted from <http://news.bbc.co.uk/1/hi/uk/6258034.stm>. © bbc.co.uk/news.

Must taxes distort?

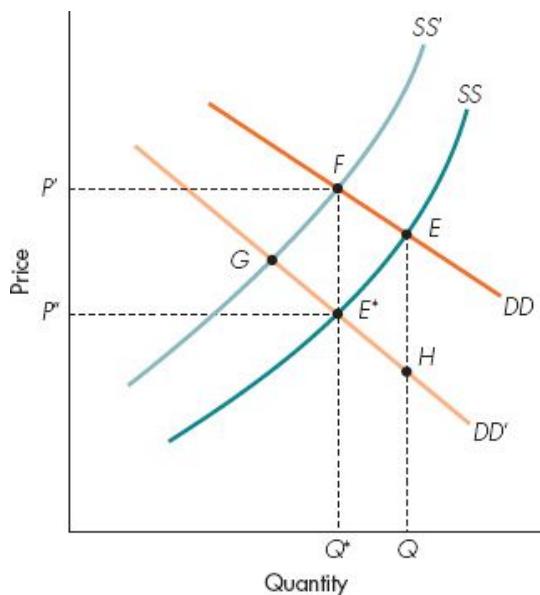
Government needs tax revenue to pay for public goods and make transfer payments. Figure 14.6 shows what happens when a tax is levied but supply is completely inelastic. There is no change in equilibrium quantity. Hence there is no distortionary triangle. The equilibrium quantity remains the efficient quantity.

We can make this into a general principle. When either the supply or the demand curve for a good or service is very inelastic, a tax leads to a small change in equilibrium quantity. Hence the deadweight loss is small. Given that the government must raise some tax revenues, waste is smallest when the goods that are most inelastic in supply or demand are taxed most heavily.

In the UK tax system, the most heavily taxed commodities are alcohol, fuel and tobacco. Alcohol, fuel and tobacco have inelastic demand.

So far, we have discussed the taxes that do least harm to efficiency. Sometimes taxes improve efficiency and reduce waste. The most important example is when externalities exist.

Cigarette smokers pollute the air for other people but ignore this in deciding how much to smoke. They cause a harmful consumption externality, as discussed in Case 14.2. Figure 14.7 shows the supply curve SS of cigarette producers. With no production externalities, SS is also the marginal social cost curve. DD is the private demand curve; that is, the marginal benefit of cigarettes to smokers. Because of the harmful consumption externality, the marginal social benefit DD' lies below DD .



Given private demand DD and supply SS , free market equilibrium is at E with a quantity Q . With a negative consumption externality, the social marginal benefit is DD' lying below DD . E^* is the socially efficient point at which output is Q^* . At this output the marginal externality is E^*F . By levying a tax of exactly E^*F per unit, the government can shift the private supply curve from SS to SS' , leading to a new equilibrium at F at which the socially efficient quantity Q^* is produced and the deadweight burden of the externality EHE is eliminated.

Figure 14.7 Taxes to offset externalities

With no tax, equilibrium is at E , but there are too many cigarettes. The efficient quantity is Q^* , which equates marginal social cost and marginal social benefit. Suppose the government levies a tax, equal to the vertical distance $E * F$, on each packet of cigarettes. With the tax-inclusive price on the vertical axis, the demand curve DD is unaffected, but the supply curve shifts up to SS' . Each point on SS' then allows producers to receive the corresponding net-of-tax price on SS .

The tax shifts equilibrium to F . The efficient quantity Q^* is produced and consumed. Consumers pay P' and producers get P' after tax is paid at the rate E^*F per packet.

The tax rate $E * F$ guides the free market to the efficient allocation. A lower tax rate (including zero) leads to too much consumption and production of cigarettes. A higher tax rate than E^*F moves consumers too far up their demand curve, causing too little consumption and production.

A tax rate E^*F leads to the efficient quantity because this is the size of the marginal externality when the efficient quantity Q^* is produced. A tax at this rate makes consumers behave as if they took account of the externality, though they think only about the tax-inclusive price.

When externalities induce distortions, the government can improve efficiency by levying taxes. The fact that alcohol and tobacco have harmful externalities is another reason to tax them heavily.

MATHS 14.1

USING A TAX TO INTERNALIZE THE NEGATIVE EXTERNALITY

Consider the same example as in Maths 13.1 and two firms, A and B. Firm A pollutes the lake used by firm B for fishing. Here, we briefly report the main features of the two firms. The cost function of firm A is $TC_A = TC_A(Q_A, P_A)$, where Q_A is the quantity produced by A and P_A is the level of pollution of firm A. The cost function of firm A has the following properties:

- It increases with the output produced: $\partial TC_A / \partial Q_A > 0$
- It decreases with the level of pollution: $\partial TC_A / \partial P_A \leq 0$

Firm B has the total cost function $TC_B = TC_B(Q_B, P_A)$, with the following properties:

$$\partial TC_B / \partial Q_B > 0 \quad \text{and} \quad \partial TC_B / \partial P_A > 0$$

If property rights are not assigned and with no government intervention, the optimal level of pollution chosen by firm A satisfies the condition:

$$-(\partial TC_A / \partial P_A) = 0$$

that is, marginal cost of pollution equals marginal revenue of pollution (zero in this case).

Firm A is polluting more than it should, so we can tax firm A in such a way that the socially efficient level of pollution is reached. Those kinds of taxes are also called Pigouvian taxes.

When firm A has to pay a tax for its polluting activity, the profit function of firm A becomes:

$$\pi_A = pQ_A - TC_A(Q_A, P_A) - tP_A$$

where t is the tax rate and p is the market price for the output of firm A. Now the optimal level of pollution (the one that maximizes firm A's profits) is:

$$-(\partial TC_A / \partial P_A) - t = 0 \Rightarrow -(\partial TC_A / \partial P_A) = t$$

The tax simply increases the marginal cost of polluting. Now the level of pollution that maximizes the profits is lower than before since $-\partial TC_A / \partial P_A$ must be equal to t , that is, greater than zero.

What tax level provides the efficient solution for the externality?

If we set $t = \partial TC_B / \partial P_A$, we obtain the efficient solution. Why? Because by setting $t = \partial TC_B / \partial P_A$, firm A now chooses a level of pollution that satisfies $-\partial TC_A / \partial P_A = \partial TC_B / \partial P_A$. In deciding the optimal level of P_A , firm A now takes into account the effect that its decision has on firm B. In practice, in order to obtain the efficient solution we need to set the tax (paid by the polluting firm) equal to the marginal social cost of pollution. This kind of tax is known as *Pigouvian tax*.

14.4 Taxation and supply-side economics

Supply-side economics analyses how taxes and other incentives affect national output when the economy is at full capacity.

Suppose the government cuts spending and tax rates. What are the effects? First, by spending less on goods and services, the government frees some resources for use by the private sector. If the private sector is more productive than the public sector, the transfer of resources may directly raise output. Whether the private sector actually uses resources more productively than the government is unclear. It seems to do many things better but some things worse.

What about the effects of lower tax rates? Figure 14.7 suggests that tax distortions cause inefficiency. Lower taxes mean a lower deadweight burden. The size of this gain depends on supply and demand elasticity. If either elasticity is small, the social gain is low.

For example, Chapter 10 argued that labour supply is fairly inelastic for those in employment, but a bit more elastic for those thinking of joining the labour force.

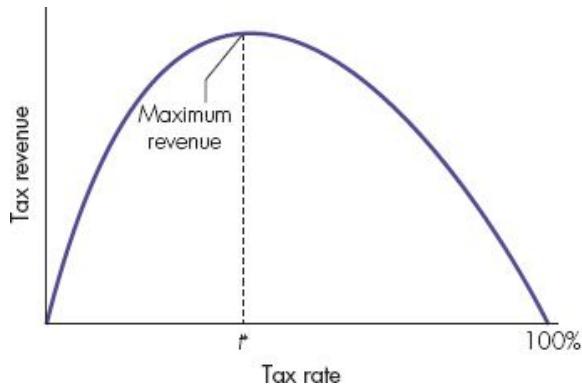
Cutting income tax rates *will* increase labour supply, but perhaps by less than many advocates of tax cuts believe.

The Laffer curve

The **Laffer curve** shows how much tax revenue is raised at each possible tax rate.

We now discuss the relationship between tax rates and tax revenues. Professor Laffer was an adviser to US President Ronald Reagan.

Figure 14.8 shows that with a zero tax rate the government gets zero revenue. At the opposite extreme, with a 100 per cent income tax rate, there is no point working and again tax revenue is zero. Beginning from a zero rate, a small increase in the tax rate yields some tax revenue. Initially tax revenue rises with the tax rate, but beyond the tax rate t^* (the tax rate at which tax revenues are maximized) higher taxes have major disincentive effects on work effort, and revenue falls.



The Laffer curve shows the relationship between tax rates and tax revenue. Moderate tax rates raise some revenue. Beyond t^* higher tax rates reduce revenue because disincentive effects greatly reduce the supply of the quantity being taxed. At a 100 per cent tax rate, supply and revenue will be zero again.

Figure 14.8 The Laffer curve

Professor Laffer's idea was that 'big government-big tax' countries were at tax rates above t^* . If so, tax cuts were the miracle cure. The government would get *more* revenue by cutting taxes. By reducing the tax distortion and increasing the amount of work *a lot*, lower tax rates would be more than offset by higher incomes to tax.

The shape of the Laffer curve is not in dispute. However, many economists disputed the view that *in fact* tax rates were above t^* . Most economists' reading of the empirical evidence is that our economies were always to the left of t^* . Cutting income tax rates may eliminate some of the deadweight burden of distortionary taxation, but governments should probably expect their tax revenue to fall if tax rates are cut.

Governments wishing to avoid borrowing need to cut their spending if they wish to cut the tax rate.

14.5 Local government

So much for central government. What about local government? Local government spends on things from sweeping the streets to providing schooling. This is financed both by local taxes and by money from central government financed by national taxes. Local government is also responsible for some types of regulation, for example land use or *zoning* laws.

Economic principles

Why don't we make central government responsible for everything? First, diversity matters. People are different and do not want to be treated the same. Civic pride is necessarily local. Second, people feel that central government is remote from their particular needs. Even if central government paid attention to local considerations, it would find it hard to do so efficiently.

We examine two important models of local government. The *Tiebout model*² emphasizes diversity. Some people want high spending, good public services and high local taxes; others want low local taxes even if this means poor public services. The Tiebout model is sometimes called the *invisible foot*: people cluster in the area providing the package of spending and taxes they want. The 'invisible foot' allocates resources efficiently via competition *between* local governments.

The 'invisible foot' is a crude incentive structure. First, it is hard to move between local authorities. You may lose your place in the queue for housing provided by that local authority. Second, if much of local authority revenue comes from central government, the levels of spending and taxes may be insensitive to the wishes of local residents.

Even if the 'invisible foot' led to efficiency, it might also lead to inequity. The rich are likely to cluster together in suburbs. Then they pass zoning laws specifying a minimum size for a house and its garden. The poor cannot move into that area. By forming an exclusive club, the rich ensure that their taxes do not go to supporting the poor. The poor get stuck with one another in inner-city areas whose governments face the biggest social needs but have the smallest local tax base.

The Tiebout model assumes that residents consume the public services provided by their own local authority. When each unit of local government has responsibility for a small geographical area, this may be a bad assumption. If a city supplies free art galleries, financed by taxes on city residents, the rich still come in from the suburbs to use these facilities. Conversely, urban trendies spend their Sundays enjoying countryside facilities supported by rural taxes. In both cases, provision of public services in one area confers a beneficial externality on nearby areas.

Economic theory suggests an answer to this problem. Widen the geographical area of each local government until it includes most of the people who use the public services it provides. It may make sense to have an integrated commuter rail service and inner-city

subway, and to subsidize it to prevent people driving through congested streets. However, only a local government embracing both the suburbs and the inner city is likely to get close to the efficient policy.

The Tiebout model favours a lot of small local government jurisdictions to maximize choice and competition between areas. However, the presence of externalities across areas suggests larger jurisdictions to ‘internalize’ externalities that would otherwise occur. The right answer may involve a bit of both.

14.6 Economic sovereignty

Nowadays, no country is an economic island, cut off from the rest of the world. We examine the world economy in Part Five, but some issues cannot be postponed until then. In a democratic country insulated from the rest of the world, the government is sovereign: while it retains democratic support and observes existing laws, it has the final say in policy design. Sometimes central government chooses to delegate powers to local government.

Economic sovereignty is the power of national governments to make decisions independently of those made by other governments.

What this account ignores is the existence of other countries. How do interactions with the rest of the world affect the sovereignty of national governments?

Even in a quarantined economy, governments cannot do anything they like. Within market economies they have to work within the forces of supply and demand. For example, in Section 14.3 we argued that it is generally more efficient to have high tax rates on things for which the demand or supply is inelastic. High tax rates on things with elastic supply and demand induce large distortions since equilibrium quantity is very sensitive to the price. We now apply this insight to economies open to interactions with the rest of the world.

International capital is now highly mobile across countries. Suppose the UK government tries to levy a large tax on capital in Britain. Lots of capital will quickly move elsewhere to escape the high taxes. The *tax base*, in this example the quantity of capital available for taxing in Britain, quickly shrinks. So the high tax *rates* may raise little tax *revenue*. In contrast, since people are much less mobile than capital across national boundaries, the tax base for taxing workers’ incomes in Britain is much less sensitive to tax rates than the tax base for capital taxes.

Even people are more mobile across national boundaries than they were a few decades ago. Communication is easier and transport costs are lower. Migration affects not just taxation but government spending as well. Suppose a country wishes to implement a generous welfare state. As a closed economy, all it has to worry about is how much of its tax base disappears from work into leisure. If welfare is too generous, people may not work enough. As an open economy, it also has to consider whether more generous

welfare provision will lead to more migration into the country as foreigners take advantage, legally or illegally, of the generous welfare provision.

Closer economic integration with other countries – through trade in goods and movement of factors of production – effectively undermines the sovereignty of nation states. If the tax rate was 80 per cent in Liverpool but 20 per cent in Manchester, one would expect big movements of capital and people from Liverpool to Manchester. The tax base in Liverpool would evaporate. The local government of Liverpool has limited local sovereignty because it is effectively in competition with Manchester.

The economic sovereignty of nation states, their freedom to do what they want, is steadily being constrained by competition from foreign countries. More than one in ten cans of beer now consumed in England was bought by British households in France, hopping across the Channel to take advantage of lower alcohol taxes in France. UK Chancellors of the Exchequer, caught between the pressure to raise revenue and support jobs in the UK drinks industry, have been cutting the real value of UK alcohol taxes. They have already lost the sovereignty to set tax rates at the high levels that they would have liked.

National sovereignty is undermined not just by competition between countries for tax bases but also by two other forces. The first is other cross-country spillovers such as acid rain, greenhouse gases or the threat of pollution from a nuclear accident. Banning nuclear power generation in southern England has limited value if northern France is studded with nuclear power stations.

The second is the scope for redistribution. Economics is about equity as well as efficiency. In an important sense, the right jurisdiction for government is the area within which citizens feel sufficient identity with one another that the rich are prepared to pay for the poor, and the fortunate are prepared to assist the unlucky. European nation states have long histories and strong national identities. But these are not always set in stone. Countries such as Belgium, Italy, Spain and the UK have faced strong internal pressures to allow parts of their country to secede. In the opposite direction, some Europeans now feel as much a citizen of Europe as of their own particular nation.

Nation states are not yet obsolete. But they are coming under pressure. Further developments in technology will increase the transnational scope of economic interactions and cultural identity. The proliferation of ecommerce and the Internet will only accelerate this process.

14.7

Political economy: how governments decide

Political economy is the study of how governments make decisions.

Firms are in business to make profits for their owners. Individuals buy affordable combinations of goods that yield them most satisfaction. These simple assumptions let economists explain most consumer and business decision making. What about government decision making?

Government is the most important single player in the economy. It is important to develop theories of how governments behave. There is no point analysing the consequences of a policy that a sane government will never implement.

Voters elect governments to set spending and taxing, pass new laws and establish new regulations. The electorate chooses among alternative policy *packages* offered by competing parties, but is rarely allowed a referendum on each issue.

The government does not simply do the bidding of society. Government has its own agenda, which may be to promote what it thinks is good for the public or simply to get re-elected.

The median voter

If everyone was identical and of one mind, public decision making would be trivial. Through the political process, society tries to reconcile different views and different interests.

Figure 14.9 shows 17 different voters and how much each wants the government to spend on the police. A dot shows each voter's preferred amount. Assume that a voter whose ideal amount is £250 will think that £300 is better than £400 if these are the only choices on offer, and will prefer £200 to £100. Each person has *single-peaked* preferences, being happier with an outcome the closer it is to his peak or preferred level.

The **median voter** on an issue is the person whose preferences are such that half the population's preferences on the issue lie on one side and half the population's preferences lie on the other side.



Each dot represents the preferred expenditure of each of 17 voters. The outcome under majority voting will be the level preferred by the median voter. Everybody to the left will prefer the median voter's position to any higher spending level. Everybody to the right will prefer it to any lower spending level. The median voter's position is the only position that cannot be outvoted against some alternative. Hence it will be chosen.

Figure 14.9 The median voter

There is a vote on how much to spend on the public good called police. A proposal to spend £0 is defeated by 16 votes to 1. Only the voter who is the left-hand dot in Figure 14.9 votes for £0 rather than £100. From either extreme, as we move to the centre more people vote for a particular proposal. With 17 voters, the median voter is the person who wants to spend the ninth-highest amount on the police. Eight voters want to spend more; eight want to spend less.

Any proposal for higher spending than the median voter's preferred amount can be defeated. The median voter, plus the eight voters below him, all vote against. But any proposal for lower spending is also defeated. The median voter, and the eight voters above him, all vote against. Hence, the median voter gets his way by majority voting.

Log-rolling

So far we have assumed each issue is voted on independently. Making decisions through legislative compromises is much more complicated when votes can be traded between different issues. Groups of politicians form parties or coalitions within which some vote trading can take place.

Log-rolling is a vote for another person's preferred outcome on one issue in order to exchange for their vote your preferred outcome on another issue.

For two issues, A and B, and three politicians, Tom, Dick and Harry, Table 14.4 shows the value of each outcome to each politician. Suppose each person votes for a proposal only if the outcome is positive. Tom votes against A and B, Dick votes against A but for B, and Harry votes for A but against B. Both issues are defeated on a majority vote.

Table 14.4 Log-rolling

Politician	Issue A	Issue B
Tom	-4	-1
Dick	-3	4
Harry	6	-1

ACTIVITY 14.1

HUNTING THE MEDIAN VOTER

After Labour lost the 1979 general election it moved to the left. This pleased party activists but took the party too far away from the preferences of the median voter. The Conservatives were in power for the next 17 years. After heavy defeat in 1983, successive Labour leaders slowly moved the party back to the middle ground that the median voter inhabits. Labour focus groups interviewed people directly to clarify the median voter's view on different issues. The result? Labour victories in 1997, 2001 and 2005.

Did Labour abandon its principles to win and keep office? It gave up old traditions of high welfare spending and high, visible taxes. But, when Gordon Brown was the Chancellor of the Exchequer, he helped the poor substantially without frightening the middle classes. As a result of his budgets, the post-tax income of the poorest 20 per cent of people rose by over 10 per cent (see Figure 14.4 again).

How did he do it? Not by raising income tax or VAT. Some of it was financed by stealth taxes, such as the tax treatment of pension funds, which the median voter did not initially notice or understand. Some was financed by making transfer payments more selective. Instead of a universal benefit, scarce resources were concentrated only on those who really needed them. Some of it was financed by economic growth: as incomes grew, given tax rates yielded more tax revenue, which was given mainly to the poor.

Unusually, Labour did not take credit for the extent to which it helped the poor. This kept the median voter sweet (the middle classes were not told repeatedly how they were paying too much to support the poor), but upset some traditional Labour supporters (who probably still voted Labour anyway).

The electoral success of the Labour Party partly ended in May 2010, when the Conservatives won the majority of seats (not the necessary number to have an overall majority though) at the general election and a coalition government between the Conservatives and the Liberal Democrats was formed. Did the median voter change his mind?

Questions

- (a) In a country with two parties, suppose both end up with almost identical policies in the centre ground. What does this tell you about (i) the ideology of the party leaders, and (ii) the extent to which party activists trade off the desire for power and their political beliefs?



David Cameron (Conservative), Nick Clegg (Liberal Democrat) and Gordon Brown (Labour) in a live televised debate in May 2010 in the lead-up to the UK general election. © Gareth Fuller/PA Archive/Press Association Images

- (b) Suppose we could order voters from left to right with equal numbers of voters holding each possible opinion. If everyone votes for the party nearest their own beliefs, where should the two parties locate to maximize their vote?
- (c) Now, however, suppose that people abstain if the party is not close to the voters' ideal positions. Does this change the optimal positioning of party manifestos? *To check your answers to these questions, go to page 679.*

Now suppose Dick and Harry vote together. They vote for A, which Harry really wants, and for B, which Dick really wants. Dick gains 4 since B passes, and loses only 3 when A passes. Harry gains 6 when A passes and loses only 1 when B passes. By forming a coalition that allows them to express the intensity of their preferences, they do better than under independent majority voting, when neither A nor B would have passed.

Many decisions in the European Union reflect log-rolling. Individual countries get favourable decisions on issues they really mind about, but are expected to repay the favour on other issues.

Commitment and credibility

Chapter 9 introduced **credibility** and **commitment** in the context of games between firms. Similar ideas apply to the political economy of policy design. Because expectations about the future affect current decisions, politicians are tempted to make optimistic promises about the future in the hope of influencing people today.

Our discussion of strategic entry deterrence in Chapter 9 gives you all the clues you need to think about political credibility. Project your imagination into the future and consider how politicians will then want to behave. Use this insight to form smart guesses today about which promises are credible and which are not.

For example, most post-war Labour governments were big spenders, which required high taxation. When out of office, Labour promises of low spending and low taxes when next in government were not very credible. Gordon Brown's Code for Fiscal Stability was an attempt to enhance Labour's credibility by openly and repeatedly committing to a tough policy that would then be politically costly to abandon. With so much political capital invested in prudence and the Code for Fiscal Stability, the government would look very stupid if it subsequently abandoned it.

A **credible** promise about future action is one that is optimal to carry out when the future arrives.

A **commitment** is a current device to restrict future room for manoeuvre to make promises more credible today.

Recently, many countries have adopted a commitment that has been very successful. They have made the central bank operationally independent of government control, as Labour did with the Bank of England in 1997. The government chooses the aim of monetary policy – to keep inflation low – but the Bank alone now decides what interest rates are needed to achieve this. By keeping the government's hands off interest rates, central bank independence removes the temptation for the government to overheat the economy in pursuit of a pre-election boom.

Policy co-ordination

Chapter 9 contained another useful insight for modern political economy. In discussing games between oligopolists, we showed that collectively they make more profit acting as a joint monopolist than by acting without co-ordination. In the language you later learned in Chapter 13, when actions are interdependent and externalities matter, the efficient solution needs to take these spillovers fully into account. Internalizing externalities means stopping free-riding.

Policy co-ordination is the decision to set policies jointly when two interdependent areas have big cross-border spillovers.

The more interdependent different nation states become, the more it may be necessary to co-ordinate national policies rather than formulate them in isolation. Global warming is one example, but many forms of regulation and taxation fall under this heading.

French tax rates on alcohol are so much lower than UK rates that UK Chancellors can no longer set UK alcohol taxes as high as they would like. The UK would like continental tax rates on alcohol to be higher. Conversely, continental Europeans complain about low levels of worker protection in the UK and the competitive edge this may give UK firms.

Pressure for closer policy co-ordination is likely to increase as globalization continues.

Summary

- Government revenues come mainly from **direct taxes** on personal incomes and company profits, **indirect taxes** on purchases of goods and services, and **contributions** to state-run social security schemes. Government spending comprises **government purchases** of goods and services and **transfer payments**.
- Governments intervene in a market economy in pursuit of distributional equity and allocative efficiency. A **progressive tax-and-transfer system** takes most from the rich and gives most to the poor. The UK system is mildly progressive. The less well off receive transfer payments and the rich pay the highest tax rates. Although some necessities, notably food, are exempt from VAT, other goods intensively consumed by the poor, notably cigarettes and alcohol, are heavily taxed.
- **Externalities** are cases of market failure where intervention may improve efficiency. By taxing or subsidizing goods that involve externalities, the government can induce the private sector to behave as if it takes account of the externality, eliminating the **deadweight burden** arising from the misallocation induced by the externality distortion.
- A **public good** is a good for which one person's consumption does not reduce the quantity available for consumption by others. Together with the impossibility of effectively excluding people from consuming it, this implies all individuals consume the same quantity, but they may get different utility if their tastes differ.
- A free market will undersupply a public good because of the **free-rider problem**. Individuals need not offer to pay for a good that they can consume if others pay for it. The socially **efficient** quantity of a public good equates the marginal social cost of production to the sum of the marginal private benefits over all people at this output

level. Individual demand curves are vertically added to get the social demand or marginal benefit curves.

- Except for taxes to offset externalities, taxes are **distortionary**. A **wedge** between the sale price and purchase price prevents the price system equating marginal costs and marginal benefits. The size of the **deadweight burden** is higher, the higher is the marginal tax rate and the size of the wedge, but also depends on supply and demand elasticities for the taxed commodity or activity. The more inelastic are supply and demand, the less the tax changes equilibrium quantity and the smaller is the deadweight burden.
- **Tax incidence** describes who ultimately pays the tax. The more inelastic is demand relative to supply, the more incidence falls on buyers not sellers.
- Rising tax rates initially increase tax revenue but eventually lead to such large falls in the equilibrium quantity of the taxed commodity or activity that revenue falls. Cutting tax rates will usually reduce the deadweight tax burden but might increase revenue if taxes were initially very high. Few economies are in this position. Lower tax rates usually reduce tax revenue.
- The **economic sovereignty** of nation states is reduced by cross-border mobility of goods, capital, workers and shoppers. Policy co-ordination may increase efficiency by making decisions reflecting previously neglected policy spillovers.
- **Political economy** examines political equilibrium and incentives to adopt particular policies.
- When all those voting have single-peaked preferences, majority voting achieves what the **median voter** wants.

Review questions



EASY

- 1 Which of the following are public goods: (a) the fire brigade, (b) clean streets, (c) refuse collection, (d) cable television, (e) social tolerance, (f) the postal service?
- 2 How would you apply the principles of horizontal and vertical equity in deciding how much to tax two people, each capable of doing the same work, but one of whom chooses to devote more time to sunbathing and therefore has a lower income?
- 3 Classify the following taxes as progressive or regressive: (a) a higher tax on luxury goods than on necessities, (b) taxes in proportion to the value of owner-occupied houses, (c) a tax on beer, (d) a tax on champagne.
- 4 There is a flat-rate 30 per cent income tax on all income over £2000. Calculate the average tax rate (tax paid divided by income) at income levels of £5000, £10 000 and £50 000. Is the tax progressive? Is it more or less progressive if the exemption is raised from £2000 to £5000?

5 Common fallacies Why are these statements wrong? (a) If the government spends all its revenue, taxes are not a burden on society as a whole. (b) Taxes always distort markets. (c) Political economy is just an excuse to waffle, and cannot be made rigorous.

MEDIUM

6 The lake of Mangrovia is polluted by a firm. A clean lake is considered a public good by the local community. Two residents, Sam and Ronald, are interested in reducing the level of pollution in the lake. The following table shows the marginal benefits (or marginal willingness to pay) of the two residents for each unit of pollution reduction:

Units of pollution reduced	<i>MB</i> Sam	<i>MB</i> Ronald
1	£30	£12
2	£20	£9
3	£15	£4
4	£5	£2

Suppose that the cost of reducing pollution is constant at £19. How much pollution would you expect to be reduced? What is the efficient amount of pollution reduction?

- 7 A firm that produces steel is polluting the air. Assume that the marginal cost of producing steel is constant at £4. The inverse market demand for steel is $P = 44 - 2Q$, where P is the price of steel and Q is the quantity of steel. The air pollution associated with steel production is creating an externality given by £2 Q . Assuming that the market for steel is competitive, what is the profit-maximizing level of steel when only marginal private costs are taken into account? The marginal social costs are given by the sum of the marginal private costs plus the externality. What is the social level of steel output? Show your solution graphically. What is the social loss associated with the externality? How can we solve this externality problem using taxation?
- 8 Why does society try to ensure that every child receives an education? Discuss the different ways this could be done and give reasons for preferring one method of providing such an education.
- 9 Hypothecation is the promise to use tax revenue from a product to achieve benefits for the group who bear the tax, for example using the London congestion charge to improve London's public transport or using tobacco taxes to build health centres for smokers. (a) Why are politicians attracted by hypothecation and (b) why are economists not attracted by hypothecation?
- 10 Suppose the local government of a city levies high taxes on its residents and does not provide them with enough public goods. What would the unsatisfied residents do according to the Tiebout model? What are the implications of the model on the city?
- 11 Tom, Dick and Harry vote for issues A and B. Their votes are given in the table below. According to the simple majority rule, which issue will pass? Is there a chance of log-rolling among the voters? If there is, which issue will pass?

Politician	Issue A	Issue B
Tom	-4	-1

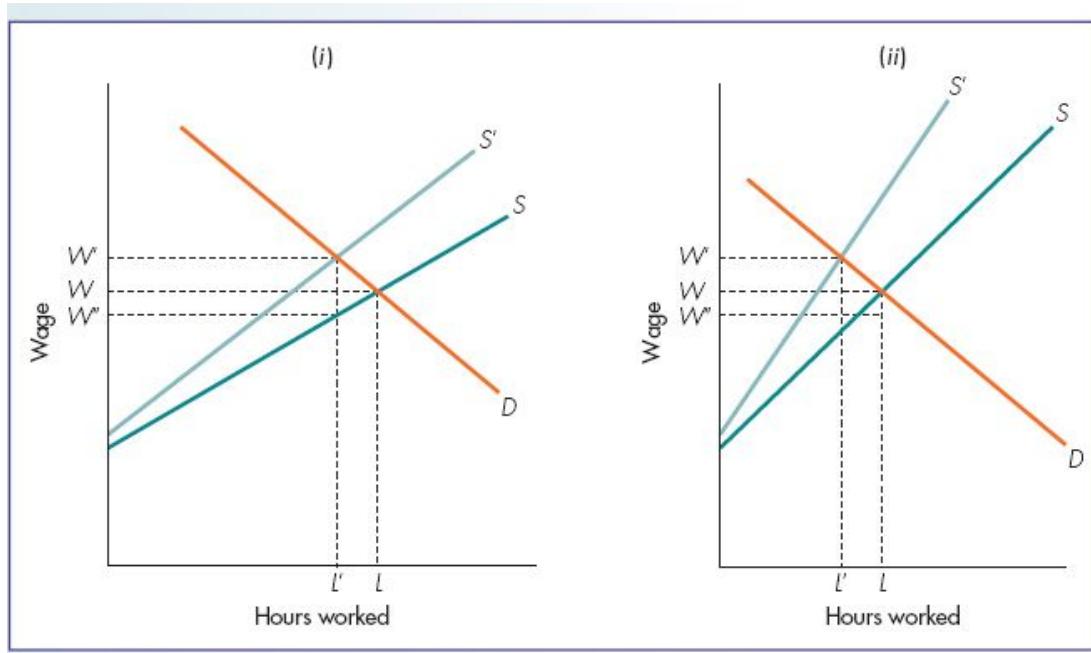
Dick	-3	4
Harry	6	-1

HARD

- |2 The market demand for milk is $Q^D = 60 - 4p$, while the market supply is $Q^S = p + p$, where p is the milk's price. Suppose that the government imposes a specific tax $t = 5$ on the suppliers of milk. Find the equilibrium in the milk market. Show your solution graphically. Calculate the total revenues generated by the tax.

What proportion of the tax revenue is paid by consumers? What proportion is paid by the suppliers? Explain graphically how those proportions may depend on the elasticity of demand.

- |3 (a) Suppose labour supply is completely inelastic. Show why there is no deadweight burden if wages are taxed. Who bears the incidence of the tax? (b) Now suppose labour supply is quite elastic. Show the area that is the deadweight burden of the tax. How much of the tax is ultimately borne by firms and how much by workers? (c) Refer to the graphs given below. In graph (i) the labour demand is elastic; in graph (ii) the labour demand is inelastic. In which graph does the tax burden fall to a greater extent on firms?



- |4 **Essay question** Imagine a new UK government, to the surprise of everyone, announces that income tax rates will rise by 15 percentage points in order to provide decent schools and hospitals. Describe the good and bad consequences. How did you decide what you meant by good and bad?

1 Does a tax always shift the supply curve? Yes, if we measure the gross price on the vertical axis. If we measure the net-of-tax price on the vertical axis, the tax shifts not the supply curve but the demand curve. In Figures 14.5 and

14.6, in terms of the net wage the demand curve shifts down until it passes through A'. The distance between A' and E still measures the tax and we reach exactly the same conclusions as before.

- 2 C. Tiebout, ‘A pure theory of local expenditures, *Journal of Political Economy* 64, no. 5 (1956): 416–424.

PART FOUR

Macroeconomics

Part Four studies the economy as an interrelated system. Output is demanded by firms, by households, by the government and by foreigners. Since interest rates and bank lending affect the demand for output, the financial sector interacts with the real economy. Price and wage adjustments help restore output to full capacity, but monetary policy and fiscal policy also play a role. Together, all this affects inflation and unemployment. Economies are increasingly open to foreign trade and foreign capital. The balance of payments records transactions with foreigners. The dynamics of the national economy also depend on the exchange rate policy pursued. By the end of Part Four, we can explain business cycles around full capacity and long-run growth in full capacity output.

Chapter 15 introduces the macroeconomy. Chapters 16–17 develop a basic model of output determination in the short run. Chapters 18–19 describe money, banking and how interest rates are set. Chapter 20 examines monetary and fiscal policy. Chapter 21 introduces aggregate supply and price adjustment. Chapters 22–23 look at inflation and unemployment. Chapters 24–26 discuss exchange rates, the balance of payments and macroeconomics within open economies.

The final two chapters of Part Four pull together our discussion of macroeconomics. Chapter 27 examines short-run business cycles and competing schools of economic thought about how the macroeconomy works. Chapter 28 discusses supply-side economics and the economics of long-run growth.

Contents

15 Introduction to macroeconomics

16 Output and aggregate demand

17 Fiscal policy and foreign trade

- 18** Money and banking
- 19** Interest rates and monetary transmission
- 20** Monetary and fiscal policy
- 21** Aggregate supply, prices and adjustment to shocks
- 22** Inflation, expectations and credibility
- 23** Unemployment
- 24** Exchange rates and the balance of payments
- 25** Open economy macroeconomics
- 26** Exchange rate regimes
- 27** Business cycles
- 28** Supply-side economics and economic growth

Understanding the financial crisis – a roadmap

Mastering macroeconomics will allow you a much deeper understanding of the financial crash and its lingering consequences. Listed below are sections dedicated to analysis of the crisis.

- | | |
|------------|--|
| Chapter 16 | Case 16.2 contrasts the current crisis with the depression of the 1930s. Case 16.3 examines how the crash has had long-term effects on investment in new capital. Case 16.4 explores how households and firms have reacted to the crash by tightening their belts and increasing saving. |
| Chapter 17 | Case 17.1 discusses how Japan is only just recovering, two decades after a financial crash, and draws lessons for Western economies. Case 17.2 introduces cyclical fluctuations in budget deficits, and therefore asks how much high deficits are the consequences of slow output growth since the crash. Section 17.5 contains an extensive discussion of the economics of budget deficits and government debt. |
| Chapter 18 | Concept 18.2 examines the collapse of the bank deposit multiplier, and contraction of bank lending. Case 18.1 provides an extensive discussion of the sub-prime crisis and its aftermath, documenting the chronology of the crisis and showing its effects on house prices and wealth. |
| Chapter 19 | Case 19.1 discusses reforms that might reduce the risk of future banking crises. Concept 19.1 explains quantitative easing. Activity 19.1 explores how transmission lags in monetary policy were exacerbated by the crisis. |
| Chapter 20 | Case 20.1 analyses Eurozone countries that developed acute debt and deficit problems after the crisis. Activity 20.1 explores how the mix of monetary and fiscal policy was altered once interest rates could not be reduced further. |
| Chapter 21 | Activity 21.1 discusses output gaps in European countries during the last 15 years, showing the dramatic output falls and slow recovery falls after the crisis. |
| Chapter 22 | Case 22.1 analyses why acute crises can give rise to deflation, and explains why this problem is difficult to solve. |
| Chapter 23 | Concept 23.2 introduces hysteresis, the case in which recession today can permanently reduce future supply. |

Chapter 24 Case 24.2 discusses currency wars, the competitive devaluations that may occur either when particular countries are in economic trouble or as part of a sustained strategy to grow at the expense of others.

Chapter 25 Case 25.1 examines the extent to which the crisis has led to pressure to try to reintroduce controls on financial flows between currencies.

Chapter 26 Case 26.1 examines global imbalances in saving and investment, and the consequences for government budgets since the crisis. Case 26.3 discusses how Mario Draghi, president of the European Central Bank, intervened to try to offset damaging consequences of the crisis in the Eurozone. Activity 26.1 explores deflation in Ireland, the textbook solution to becoming uncompetitive within the Eurozone.

Chapter 27 Section 27.4 discusses supply-side effects of the financial crash.

CHAPTER 15

Introduction to macroeconomics

Learning outcomes

By the end of this chapter, you should be able to:

- 1 view macroeconomics as the study of the whole economy
- 2 discuss the scope of macroeconomic analysis
- 3 show how the national accounts measure macroeconomic variables
- 4 explain the circular flow between households and firms
- 5 appreciate why leakages always equal injections
- 6 analyse comprehensive measures of national income and output
- 7 discuss whether national output contributes to national happiness

We now turn to the big economic issues, such as unemployment, inflation, economic growth and financial crashes. Macroeconomics sacrifices details to study the big picture. Macroeconomics is the study of the economy as a system.

The distinction between microeconomics and **macroeconomics** is more than the difference between economics in the small and economics in the large. The purpose of the analysis is also different.

Macroeconomics is the study of the economy as a system.

A model deliberately simplifies in order to focus on the key elements of a problem and think about them clearly. We could study the whole economy by piecing together a microeconomic analysis of every market, but it would be hard to keep track of all the economic forces at work. Our brains do not have a big enough Intel chip to make sense of it.

Microeconomics and macroeconomics take different approaches to keep the analysis manageable. Microeconomics stresses a detailed understanding of particular markets. To achieve this detail, many interactions with other markets are suppressed. In saying a tax on cars reduces the equilibrium quantity of cars, we ignore what the government does with the tax revenue. If government debt is reduced, interest rates may fall, making households more willing to borrow to buy new cars.

Microeconomics is like looking at a horse race through a pair of binoculars. It is great for details, but sometimes we get a clearer picture of the whole race by using the naked eye.

Because macroeconomics studies the interaction of different parts of the economy, it uses a different simplification to keep the analysis manageable. Macroeconomics simplifies the building blocks in order to focus on how they fit together and influence one another.

Macroeconomics stresses broad aggregates, such as the total demand for goods by households or total spending on machinery and building by firms. Like watching a horse race with the naked eye, individual details are more blurred but our full attention is on the big picture. We are more likely to notice the horse sneaking up on the rails.

The scope of macroeconomics is therefore to understand the interrelationship of the big issues that affect the economy – growth, inflation, unemployment, fluctuations and crises.

15.1 The scope of macroeconomics: the big issues

Real **gross domestic product (GDP)** measures the output of goods and services produced by an economy.

The **business cycle** refers to swings in GDP around an economy's trend rate of output growth.

Economic growth is a rise in real GDP.

The **labour force** is the number of people at work or looking for work.

The **unemployment rate** is the fraction of the labour force without a job.

The **inflation rate** is the percentage annual increase in the average price of goods and services.

What determines the total output of a country, which we call its real **GDP**? Why was there a long boom in output and house prices before 2007 but then the worst post-war crash in asset prices and output levels? Are there inevitably **business cycles**, or can output grow smoothly over time? Why do some countries **grow** faster than others over sustained periods? Will growth go on forever? After centuries asleep, why did the Chinese and Indian economic giants finally awaken?

We care not just about the output of goods and services, but also the market for **labour**. Why did unemployment rise in the 1970s but fall substantially thereafter? How much **unemployment** was caused by the financial crisis? Do workers price themselves out of jobs by greedy wage claims? Does technical progress destroy jobs? Can the government create more jobs? These are questions we need to answer in Part Four.

A third big theme is **inflation**. The price level is a weighted average of the prices of goods and services. Inflation is a rise in the price level. What causes inflation? Money growth, oil price rises or a budget deficit? Have we learned how to defeat inflation? Could a boom in China cause inflation in Europe? Will the vast money creation deliberately undertaken since 2009 to mitigate the recession lead to subsequent inflation?

With prices actually falling in some countries because of the severity of the recession, could economies experience a death spiral of falling prices?

A fourth big theme is boom and bust. Why can't we arrange for economic growth to be smooth? What caused the spectacular crash of 2007–09? Can we prevent another one? Can we even escape the previous one? How does the overhang of inherited debt now affect the Gross debt is total liabilities owed to creditors. Net debt is total liabilities minus total assets that could be sold in order to raise money to pay creditors.choices of households, firms and governments?

Which debt measure is being used makes a difference. By 2012 Japan's government **gross debt** was 220 per cent of its national output, but its **net debt** was only 140 per cent. You will see references to both in the media, and need to be clear which is being used. Typically, government net debt subtracts from gross debt only that part of assets that could be sold fairly easily. Japan would find it hard to sell its nuclear power stations.

Almost every day, the media discusses inflation, unemployment, economic growth, output fluctuations and debt. These issues help determine elections, and make people interested in macroeconomics.

15.2 Some facts and economic history

We begin with some facts. Table 15.1 puts recent performance in perspective, showing data since 1960 for annual averages for economic growth, unemployment and inflation.

The 1960s was a golden age of low unemployment, rapid growth and low inflation in advanced economies. In the early 1970s, with the world economy booming, OPEC quadrupled the price of oil. The rest of the 1970s saw high inflation, low growth and rising unemployment. After another oil price hike in 1979–80, the 1980s were another tough period. By the 1990s, inflation was coming down in most countries, and unemployment was falling in the UK and the US, though not in continental Europe.

The long period of economic success during 1990–2008 -steady growth, low unemployment and low inflation-made people confident that successful economic policies had finally been discovered. Confidence bred overconfidence, large borrowing, spiralling house prices and rash lending by banks. In 2008 the financial crash finally came, leading governments to step in to rescue failing private banks, but at the cost of a sharp rise in government debt. Table 15.1 shows that, as a result of the crisis, economic growth stalled and unemployment rates were sharply higher.

The largest emerging economies were often slow to get going, in many cases shackled by extensive government regulation and planning. Deregulation in the 1980s and 1990s had a dramatic effect. The BRICs – Brazil, Russia, India and China – took off. Table 15.1 shows the dramatic rise of China, which is now approaching 40 years of 10 per cent annual real growth. Although the BRICs are growing quickly, they begin from such low levels that income per person is still well below that enjoyed in advanced economies.

Table 15.1 The big picture, 1960–2013 (annual averages, %)

	1960s	1970s	1980s	1990s	2000–07	2008–13
Real GDP growth						
UK	2.9	2.4	2.4	2.1	3.1	-0.1
US	4.3	3.3	3.1	3.1	2.6	1.0
EU	5.8	3.8	2.3	2.2	2.1	0.0
China	3.0	7.4	9.8	10.0	10.8	8.9
Unemployment						
UK	2.2	4.5	10.2	8.1	5.2	8.3
US	4.1	6.1	7.3	5.8	4.9	8.2
EU	2.5	4.0	9.3	10.5	8.3	9.8
China	-	-	2.6	2.8	3.9	4.0
Inflation						
UK	3.8	12.6	7.4	3.7	1.5	3.1
US	2.4	7.1	5.6	3.0	2.7	2.4
EU	3.7	9.5	6.5	2.9	2.3	2.1
China	-	-	7.5	7.8	1.5	3.2

Sources: OECD; IMF.

Figure 15.1 takes a longer look at UK inflation, which soared in the 1970s. The Thatcher government reduced inflation after 1980, but lost control in the late 1980s when it let the economy grow too rapidly, leading to more inflation. Subsequent governments gradually got the UK back on an even keel, not least by giving the Bank of England much more independence in decisions on monetary policy.

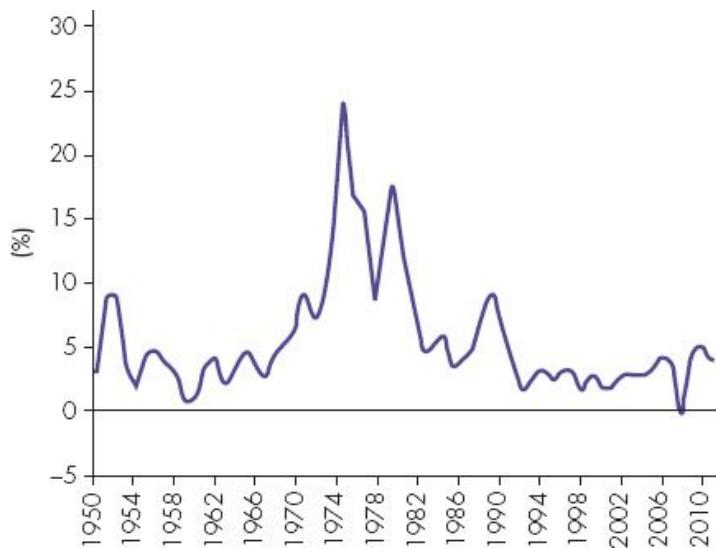


Figure 15.1 Annual UK inflation, 1950–2012

Sources: ONS.

The financial crisis temporarily reduced inflationary pressures after 2008. Despite ongoing stagnation of output growth, the UK has subsequently imported inflation (higher food and energy prices that are being bid up in the world economy by countries such as China, which continue to grow strongly).

Table 15.2 Government net debt (% of GDP)

	2007	2013
Italy	87	96
Greece	83	141
Japan	81	142
US	48	88
Germany	43	56
France	36	68
UK	28	78
Spain	18	57
Ireland	0	87

Source: Based on data from Economic Outlook Statistical Annex, © OECD, 2013, <http://www.oecd.org/eco/economicoutlook.htm>, accessed on 18/06/2013.

Table 15.2 shows the impact of the financial crisis on government debt. We examine debt relative to GDP, since the latter gives an idea of the likely tax revenues that can be used to pay interest on the debt. A rise in the debt/GDP ratio means that a country faces a greater debt burden. We show data for 2007 just before the crisis, and for 2013.

Table 15.2 shows that different countries entered the financial crisis with very different levels of government debt. Italy, Greece and Japan were already very indebted; the UK, Spain and Ireland had low debt levels prior to the crisis.

Government finances deteriorated, both because governments had to bail out private banks (the alternative would have been even worse) and because, as economies stagnated, governments got less tax revenue and had to spend more on welfare benefits. Both effects increased government debt. Countries, such as the UK, with a large financial sector experienced a large increase in government debt. Financial markets now worry that countries, such as Greece, with the largest debts will not be able to meet the interest payments on them. By the end of Part Four, we need to explain what options such countries now have.

15.3 The circular flow

The **circular flow** shows how both real resources and financial payments flow between firms and households.

The economy comprises millions of individual economic units: households, firms and the departments of central and local government. Together, their individual decisions determine the economy's total spending, income and output.

Initially, we ignore the government and other countries, leaving just firms and households. Table 15.3 shows transactions between these two sectors. Households own the factors of production (inputs to production). Households rent labour to firms in exchange for wages. Households are also the ultimate owners of firms and get their profits. Capital and land, even if held by firms, are ultimately owned by households.

Households supply inputs to firms, which use these to make output. The second row of Table 15.3 shows the corresponding payments. Households earn incomes (wages, rents, profits), which are payments by firms for using these inputs. The third row shows that households spend their incomes buying the output of firms, giving firms the money to pay for renting production inputs.

This suggests, correctly, that there are three equivalent ways in which to measure the total economic activity in an economy: (a) the value of all goods and services produced, (b) the total value of earnings arising from the factor services supplied or (c) the total value of spending on goods and services. All payments are the counterparts of real resources. For the moment, we assume all payments are spent buying real resources. We get the same estimate of total economic activity whether we use the value of production, the level of factor incomes or total expenditure on goods and services.

Table 15.3 Government net debt (% of GDP)

Households	Firms
Supply inputs to firms	Use inputs to make output
Receive incomes from firms	Rent inputs from households
Buy output of firms	Sell output to households

Household income equals household spending if all income is spent. The value of output equals total spending on goods and services if all goods are sold. The value of output also equals the value of household incomes. Since profits are residually defined – the value of sales minus the rental of inputs – and since profits accrue to the households that own firms, household incomes (from renting out inputs or from profits) equal the value of output.

Our model is still very simple. What happens if firms do not sell all their output? What happens if firms sell output not to households but to other firms? What happens if households do not spend all their incomes? The next section answers these questions. Our conclusion will be unchanged: the level of economic activity can be measured by valuing total spending, total output or total earnings. All three methods give the same answer.

Our framework still omits key features of the real world: saving and investment, government spending and taxes, transactions between firms and with the rest of the world. These are all easily remedied, later in the chapter, to create a comprehensive system of national accounts.

15.4 National income accounting

Measuring national income and output

Gross national product (GNP), also called **gross national income (GNI)**, is the total income of a country.

Gross domestic product (GDP) measures the value of output produced in a country, no matter whose citizens contribute to this production. **Gross national product (GNP)**, sometimes called **gross national income**, measures the value of the income that its citizens earn, from whatever countries this income is derived.

Value added is the increase in the value of goods as a result of the production process.

Final goods are purchased by the ultimate user, either households buying consumer goods or firms buying capital goods such as machinery.

Intermediate goods are partly finished goods that form inputs to a subsequent production process that then uses them up.

Thus, a German working in a bank in London contributes to UK GDP because that is the location at which the inputs were supplied and output produced. However, if the German then sends some of this income back to relatives in Germany, this act adds to German GNP (but not to UK GNP). Similarly, Irish foreign investments in the UK contribute to UK output and GDP, but the income derived will ultimately add to Irish GNP rather than UK GNP. Thus, GNP measures the total worldwide income of citizens of a country; GDP measures the output produced within a country, no matter which citizens produce it.

Initially we discuss a *closed economy*, not linked to the rest of the world, in which output and income are the same. First, we extend the simple circular flow between firms and households shown in Table 15.3. Transactions do not take place exclusively between a single firm and a single household. Firms hire labour services from households but buy raw materials and machinery from *other* firms. To avoid double counting, we use **value added**.

Value added is gross output minus the value of input goods used up in making that output. Closely related is the distinction between final goods and intermediate goods. Ice cream is a **final good** consumed by its ultimate user. Steel is an **intermediate good**, made by one firm but then used as an input by another firm. Capital goods are final goods because they are *not* used up in subsequent production. They do not fully depreciate during the production period under study.

An example will clarify these concepts. Suppose there are four firms in the economy: a steel maker, a producer of capital goods (machines) for the car industry, a tyre maker and

a car producer who sells to the final user, households. Table 15.4 calculates GDP for this simple economy.

Table 15.4 Calculating GDP

(1) Good	(2) Seller	(3) Buyer	(4) Transaction value	(5) Value added	(6) Spending on final goods	(7) Household earnings
Steel	Steel maker	Machine maker	£1000	£1000	-	£1000
Steel	Steel maker	Car maker	£3000	£3000	-	£3000
Machine	Machine maker	Car maker	£2000	£1000	£2000	£1000
Cars	Car maker	Households	£5000	£2000	£5000	£2000
Total transactions			£11 000			
GDP				£7000	£7000	£7000

The steel firm makes £4000 worth of steel, one-quarter sold to the machine maker and three-quarters sold to the car maker. All £4000 is value added or net output of the steel firm, paid out as household income (wages, rents, residual profits). Hence, the first two rows of the last column also add up to £4000. Firms have spent £4000 buying this steel output, but it is not expenditure on final goods. Steel is an intermediate good, used up in later stages of the production process.

The machine maker buys £1000 of steel input, converting it into a machine sold to the car maker for £2000. The value added by the machine maker is £2000 minus £1000 spent on steel input. Net revenue of £1000 accrues directly to households as income or profit. Since the car firm intends to keep the machine in the future, the full value of £2000 is shown under ‘final expenditure’ during the period.

The car producer spends £3000 on steel, used up during the period in which cars are made. We subtract £3000 from the car output of £5000 to get the value added of the car maker. Value added and household income are £2000.

Finally, the car producer sells the car for £5000 to the final consumer – households. Only then does the car become a final good. Its full price of £5000 is final expenditure.

Table 15.4 shows that the gross value of all the transactions is £11 000. This overstates the value of the goods the economy has actually produced. The £3000 that the steel producer earned by selling steel to the car producer is already included in the final value of car output. National output cannot count this twice.

Column (5) shows the value added at each stage in the production process; £7000 is the true net output of the economy. Since each firm pays the corresponding net revenue to households either as direct factor payments or indirectly as profits, household earnings are £7000 in the last column of the table. If we add up payments made to households as income and profits, we get the same measure of GDP.

Table 15.4 confirms that we also get the same answer if we measure spending on *final* goods and services. In this case, final users are households buying cars and the car producer buying the (everlasting) machinery used to make cars.

Investment and saving

Investment is the purchase of new capital goods by firms.

Saving is the part of income not spent buying goods and services.

If total output and household incomes are each £7000, but households spend only £5000 on cars, what do they do with the rest of their incomes? And who does the rest of the spending? To resolve these issues, we need **investment** and **saving**.

Households spend £5000 on cars. Their income is £7000, so they save £2000. The car maker spends £2000 on investment, buying new machinery. Figure 15.2 shows what happens to the circular flow. The bottom half of the figure shows that incomes and factor services are each £7000. But £2000 leaks out from the circular flow when households save. Only £5000 finds its way back to firms as household spending on cars.

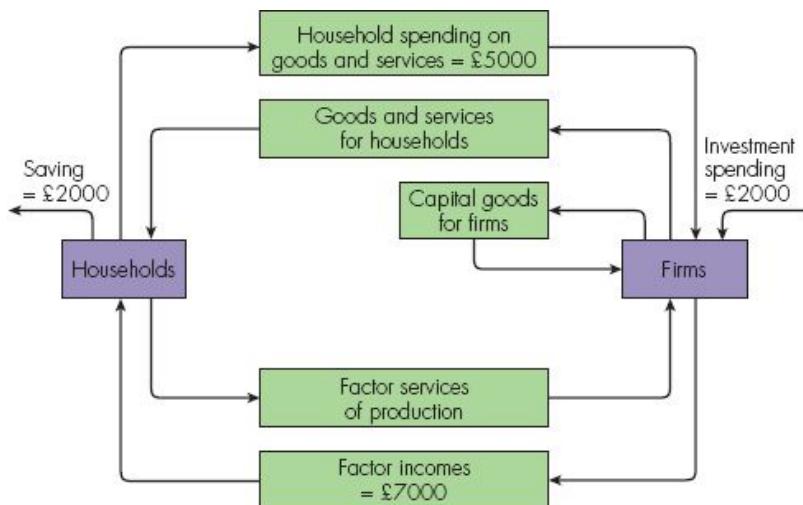


Figure 15.2 Investment, saving and the circular flow

The inner loop continues to show flows of real resources between firms and households. Firms use factor services supplied by households to make consumer goods and services for households and new capital goods for other firms. The outer loop continues to show payment flows. Factor incomes are either saved by households or spent on consumer goods. Firms also earn revenue from expenditure on new capital goods by other firms. Thus, saving is a leakage from the circular flow of payments, investment an injection into the circular flow.

The top half of the figure shows that £5000 is both the value of output of consumer goods and of household spending on these goods. Since GDP is £7000, the other £2000 comes from spending by firms themselves. It is the £2000 of investment expenditure made by the car producer buying machinery for car production.

A **leakage** from the circular flow is money not recycled from households to firms.

An **injection** is money that flows to firms without being recycled through households.

On the inner loop, firms make an output of £5000 for consumption by households and an output of £2000 of capital goods for investment by firms. On the outer loop, which relates to money payments, saving is a **leakage** of £2000 from the circular flow and investment spending is an **injection** of £2000 to the circular flow.

Is it coincidental that household savings of £2000 exactly equal investment expenditure of £2000 by firms? If not, how is the money saved by households transferred to firms to allow them to pay for investment spending?

Suppose Y denotes GDP, which also equals the value of household incomes, C denotes household spending on consumption and S denotes saving. By definition, saving is unspent income, so $Y \equiv C + S$, where the symbol \equiv means ‘is identically equal to, as a matter of definition’. Since one definition of GDP is the sum of final expenditure, $Y \equiv C + I$. Putting these two definitions together,

$$S \equiv I$$

since both are identical to $(Y - C)$.

It is no accident that saving and investment are each £2000 in our example. Saving and investment are always equal, in the absence of government and foreign sectors.

It is initially hard to grasp the difference between an equals sign and an identity sign. The latter means always equal, no matter what the values of the other variables. The equals sign means that some particular values of the variables (output, consumption, and so on) are needed in order to ensure the equality of the two sides of the equation.

Look again at the outer loop of Figure 15.2. All household spending in the top half of the figure returns to households as income in the bottom half of the figure. Investment spending by firms is matched by an income flow to households in excess of their consumer spending. Since saving is *defined* as the excess of income over consumption, investment and savings must always be equal.

These accounting identities follow from our definitions of investment, saving and income. *Actual* saving must equal *actual* investment. This need not mean *desired* saving equals *desired* investment. To study that, we need models of desired saving and investment, a task we begin in the next chapter.¹ In a market economy, financial institutions and financial markets channel household saving to the firms that wish to borrow to invest in new capital goods.

Inventories or **stocks** are goods currently held by a firm for future production or sale.

What happens if firms cannot sell all the output that they produce? Surely this creates a gap between the output and expenditure measures of GDP? Producers then add to their **inventories** or **stocks** of finished goods awaiting sale to consumers in the next period.

Stocks are sometimes called *working capital*. Not used up in production and sale during the current period, stocks are classified as capital goods. Adding to stocks is investment in working capital. When stocks are depleted, we treat this as negative investment, or disinvestment. This keeps the national accounts straight. Any unsold output is treated as temporary investment in inventories. Depletion of inventories in subsequent periods will then be treated as negative investment during the periods in which it occurs.

The domestic government and foreign countries

Firms and households are not the only sectors in the economy. We now recognize the existence of governments and of foreign countries.

Governments raise revenue both through direct taxes on income (wages, rents, interest and profits) and through indirect taxes or expenditures taxes (VAT, petrol duties and cigarette taxes). Taxes finance two kinds of expenditure. Government spending on goods and services G is purchases by the government of physical goods and services. It includes the wages of civil servants and soldiers, the purchase of computers, tanks and military aircraft, and investment in roads and hospitals.

Transfer payments are monetary payments that require no goods or services in return.

Governments also spend money on **transfer payments** or benefits, B . These include pensions, unemployment benefit and subsidies to firms.

Transfer payments do not affect national income or national output. They are not included in GDP. There is no corresponding net physical output. Taxes and transfer payments merely redistribute existing income and spending power away from people being taxed and towards people being subsidized. In contrast, spending G on goods and services produces net output, and gives rise to factor earnings in the firms supplying this output and also to additional spending power of the households receiving this income. Hence government spending G on goods and services is part of GDP. It is final expenditure since government is now an additional end user of the output.

GDP at market prices measures domestic output inclusive of indirect taxes on goods and services.

National income accounts aim to provide a logically coherent set of definitions and measures of national output. However, taxes drive a wedge between the price the purchaser pays and the price the seller receives. We can choose to value national output either at **market prices** inclusive of indirect taxes on goods and services (the price consumers pay), or at the prices received by producers after indirect taxes have been paid.

So far we have studied a closed economy not transacting with the rest of the world. We now examine an *open economy* that deals with other countries.

Households, firms and the government may buy imports Z that are not part of domestic output and do not give rise to domestic factor incomes. These goods are not in the output measure of GDP, the *value added* by domestic producers. However, imports show up in final expenditure. There are two solutions to this problem.

MATHS 15.1

THE CIRCULAR FLOW REVISITED

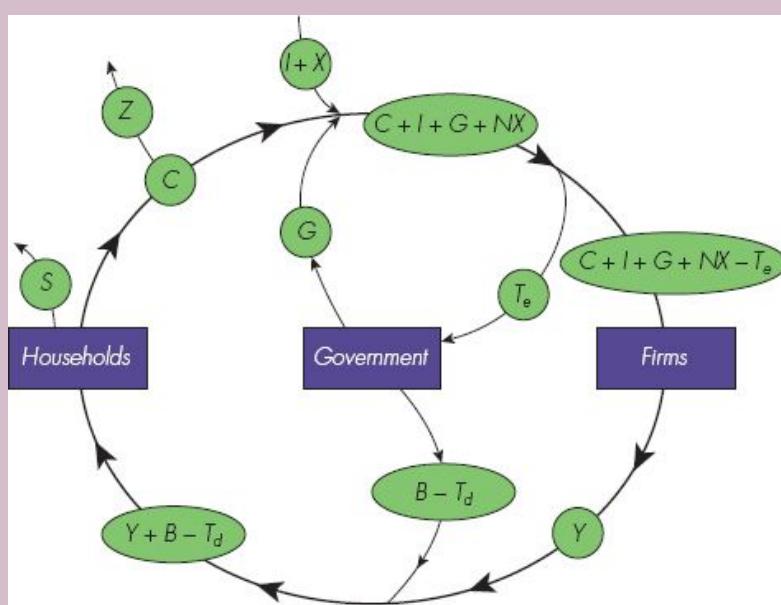
If Y is GDP at basic prices, the economy's value added or net output is goods and services for domestic consumption, investment, government spending and net exports minus indirect taxes:

$$Y \equiv \text{GDP at basic prices} \equiv C + I + G + NX - T_e$$

Household incomes at basic prices are supplemented by welfare benefits B less direct taxes T_d . This gives us *personal disposable income* ($Y + B - T_d$) available for households to spend. Suppose for the moment that saving is done only by households. Disposable income must be spent on consumption or saving:

$$Y + B - T_d \equiv C + S$$

The figure below shows the extended picture of the circular flow.



The figure extends the circular flow between households and firms to include the government and foreign sectors. Firms make factor payments Y to households. Disposable income $Y + B - T_d$ also includes transfer payments B less direct taxes T_d . Disposable income goes on saving S or consumption C . This spending is augmented by injections of government spending G on goods and services and by investment spending I and by exports X , but is reduced by the additional leakage Z .

into imports. From $C + I + G + NX$ or GDP at market prices, we must subtract the leakage of indirect taxes T_e to get GDP at basic prices Y which firms pay out to households.

Round the top loop of the figure, consumption C at market prices is supplemented by injections of investment spending I , net exports NX and government spending G . From this GDP at market prices, we subtract indirect taxes T_e to get GDP at basic prices.

$$S + NT + Z \equiv I + G + X$$

where net taxes NT are direct and indirect taxes minus welfare benefits. Investment, government spending and exports are all injections to the circular flow that do not originate from households. Conversely, household spending leaks out, directly or indirectly, through saving, taxes (net of benefits) and imports: only the remaining spending flows back to domestic firms and round again as household incomes. If there is no government or foreign sector, this becomes $S \equiv I$, as we had before. More generally,

$$(S - I) + (NT + G) \equiv (X - Z)$$

Exports minus imports measure the country's net external surplus achieved by exporting more than it imports. This external surplus allows some domestic combination of a private sector surplus (saving minus investment) or a government surplus (net taxes minus government spending).

We could subtract the import component separately from C, I, G and X and measure only final expenditure on the domestically made bit of consumption, investment, government spending and exports. But it is easier to continue to measure total final expenditure on C, I, G and **exports** (X) and then to subtract from this total expenditure on **imports** (Z). It comes to exactly the same thing.

Exports (X) are domestically produced but sold abroad.

Imports (Z) are produced abroad but purchased for use in the domestic economy.

In the previous section, we saw that our definitions should imply that total income, expenditure and output measures of total activity should coincide. We now explain how this works once we introduce the government and foreign sectors as well. The complete system of national accounts is summarized in Figure 15.3.

GNP at market prices (also GNI at market prices)	Net property income from abroad	Net property income from abroad	Depreciation	Indirect taxes	National income (NI) = NNP at basic prices	Rental income		
	G	GDP at market prices	NNP at market prices			Profits		
	I					Income from self-employment		
	NX							
	C					Wages and salaries		
COMPOSITION OF SPENDING ON GNP		DEFINITION OF GDP		DEFINITION OF NNP		DEFINITION OF NATIONAL INCOME		
						FACTOR EARNINGS		

Figure 15.3 Summarizing the national accounts

We begin on the left with gross national product (or gross national income – same thing) at market prices. The second column is the expenditure measure of GNP, which comprises spending by households on consumption, spending by firms on investment, spending by government goods and services (which we could think of as government contributions to consumption and investment), net exports (the excess of exports over imports) and, finally, net international transfers from abroad.

This last item is sometimes called *net international property income*, since most transfers arise from the return on assets held abroad (minus the return paid by us to foreigners holding assets in our country). International transfer payments also include outflows in aid by the UK when an earthquake hits Haiti, and inflows from remittances of cash to UK workers temporarily working abroad.

The third column takes us from GNP to GDP; that is, gross output during the period. Whereas net international transfer payments add to our income, they do not add to our physical output of goods and services. Hence, Figure 15.3 deducts these from GNP to get to GDP.

The fourth column shows the difference between gross and net output. Net means deducting depreciation of physical capital – buildings and machinery wear out or become obsolete. Statisticians make some fairly heroic guesses sometimes about how much depreciation is going on. So far, our national accounts leave out depreciation of environmental capital. One day, if global warming continues, it will become essential to make explicit estimates for depreciation of our environmental capital.

GDP at basic prices measures domestic output exclusive of indirect taxes on goods and services.

The fifth column shows the role of indirect taxes. Measurements at market prices include indirect taxes in the statistics; this reflects prices paid by consumers. Measurements at **basic prices** remove the effects of indirect taxes to reflect the prices received by producers.

The final column shows the net incomes that accrue to the inputs (factors of production) used in making output. Total factor incomes arise from the supply of labour to earn wages and salaries, self-employed work, the supply of land to earn rent or the supply of capital to earn profits.

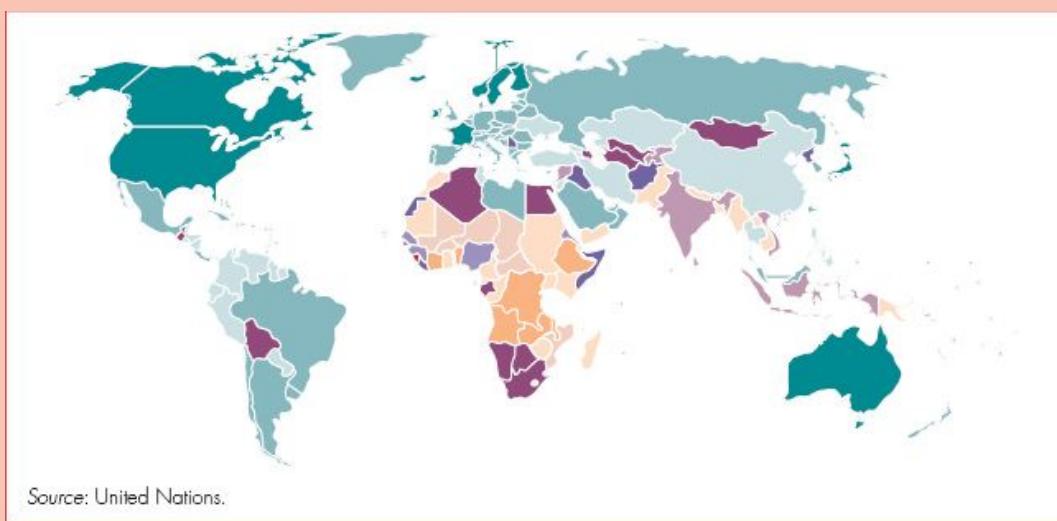
Understanding Figure 15.3 is a key step in mastering the definitions, confirming that they make sense, and checking that we have not left anything out.

CASE 15.1

PROBLEMS IN MEASURING GDP

When things are traded in a market, or embedded in government tax statistics, they are relatively easy to measure. Many of our measurement difficulties arise precisely because some of the most valuable things are not easily measurable. GDP easily captures the output of washing machines, but not of happiness, health or environmental depreciation. Since we do not buy and sell clean air or moderate temperatures in a marketplace, governments are not automatically collecting statistics for use in national accounts data.

The United Nations Human Development Index systematically tries to measure three broad dimensions of economic development – health, education and material standard of living – and produces annual statistics for all UN member countries. The map below shows the geographic range of outcomes – no prizes for guessing which colours represent prosperity and which represent poverty as measured by the Human Development Index.



Health is crudely captured by life expectancy at birth, education by the proportion of the children enrolled at school and by the proportion of adults who can read, and material standard of living by per capita GDP.

Some of these indicators are more stable than others. For example, before the financial crash, Iceland came top in the world in the UN measure, and Sierra Leone bottom. But Iceland's banks experienced the biggest crash of all, and the Icelandic economy got into serious trouble. This did not immediately affect its adult literacy or the life expectancy of its population, but these will gradually suffer unless economic prosperity can be restored.

Like sausages, economic statistics simply reflect what you put into them. If you care about democracy, equality or environmental sustainability, don't get hung up merely because your country is not doing well on the particular things that GDP does measure.

People who visit France quickly learn that the French have a good quality of life, better than you would expect simply by looking at their GDP. They enjoy a nice climate, long lunches, access to Mediterranean beaches and little congestion since they have plenty of land in relation to their population. They also retire at a relatively young age and, having long life expectancy, spend plenty of happy years in retirement. Their GDP statistics are measuring production of Renault and Peugeot, and of luxuries from Louis Vuitton and Hermès, but omit plentiful leisure, lack of stress and little congestion.

Similarly, the output of the police, civil service and teachers in free state schools is not charged for in the market and hence not automatically valued by the market. How do we measure the output of the police? Typically, national income statisticians measure the inputs (the wage bill of police forces, rent of police stations, the cost of using police cars and police computers). This is a large step in the right direction but it is far from perfect. If society becomes more unlawful, we end up choosing to have more police to counter crime. So GDP rises because we are spending more on the police force. But in reality, people are feeling less happy with the greater prevalence of crime, and resent having to 'waste' more resources on additional policing in order to counter the crime wave. Conversely, when we cut back the size of the army, GDP falls since less is being spent on the military, but we are actually receiving less defence as a consequence.

Think of all this as a health warning on GNP and GDP statistics. They measure what they measure. Unless and until electorates want to spend a lot more money collecting more comprehensive statistics, GNP and GDP will use data already being collected annually for other purposes such as taxation.

Measuring UK GDP in 2011

We now know that there are three equivalent ways in which to measure GDP, using income, output and expenditure. How does this work out in practice? Can the statisticians reconcile the different estimates? Table 15.5 shows the answer.

The first two columns show the income measure of UK GDP. The largest component is wages and salaries, then operating surpluses and, finally, other incomes. Adding these factor incomes, we get £1327 000 million, what nowadays we call £1327 billion as the measure of GDP (or gross value added, GVA) at basic prices. Adding on indirect taxes takes us to £1522 billion. A small statistical discrepancy in the income approach reduces this to a final estimate of £1519 billion at market prices.

The middle two columns reach the same answer, using the expenditure approach. Consumption spending by households is the largest component (£977 billion); then we add private investment I , the two parts of government spending on goods and services – that related to purchase of consumption goods and services and that to investment goods – and net exports to reach a total of £1517 billion. A small statistical discrepancy of £2 billion reconciles this with the estimate of GDP obtained from the income method.

The final two columns repeat the exercise, starting from estimates of value added in each industry. Nowadays, agriculture, forestry and fisheries are a tiny part of the economy. Even manufacturing is only about 10 per cent of GDP. Total production of goods, plus the output of agriculture, forestry and fisheries, still comes only to £162 billion – that is, about 11 per cent of total GDP. Construction accounts for another £103 billion, but the vast majority – £1170 billion, or 77 cent of GDP – is made up of services, everything from wholesale and retail distribution, transport and communications, financial services, government services and other privately supplied services. Adding the entire production of goods and services, we get £1519 billion as the output measure of GDP. All three measures lead essentially to the same answer. Were everything perfectly measured, the three measures would be identical even without any statistical adjustment.

Table 15.5 UK GDP, 2011, at market prices (£ billion, at 2009 prices)

Income method	£ bn	Expenditure method	£ bn	Output method	£ bn
Wages and salaries	814	Household consumption C	977	Agriculture, forestry, fisheries	9
Operating surpluses (profit, rent, interest of private and public firms)	329	Government consumption G_c	337	Manufacturing	159
Other incomes	184	Private investment I + government investment G_I	227	Other production (mining, energy, water)	78
GDP at basic prices	1327	Exports	493	Construction	103
Net indirect taxes	195	Less imports	-517	Services, of which:	1170
				Distribution, catering	213
				Transport, communications	161
Subtotal	1522	Subtotal	1517	Business/financial services	442
Statistical discrepancy	-3	Statistical discrepancy	+ 2	Govt and other services	354
GDP at market prices	1519	GDP at market prices	1519	GDP at market prices	1519

15.5 What GDP measures

A firm's accounts show how the company is doing. Our national income accounts let us assess how the economy is doing. Just as a firm's accounts may conceal as much as they reveal, we must interpret the national income accounts with care.

We focus on GDP as a measure of economic performance. Since depreciation is rather difficult to measure, and consequently may be treated differently in different countries or during different time periods, using GDP avoids the need to argue about depreciation.

Nominal GDP measures GDP at the prices prevailing when output was produced

Real GDP, or GDP at constant prices, adjusts for inflation, measuring GDP in different years at the prices prevailing at a particular date, known as the base year.

In this section we make three points. First, we recall the distinction between nominal and real variables. Second, we show how per capita GDP can provide a more accurate picture of the standard of living of an average person in an economy. Finally, we discuss the incompleteness of GDP as a measure of the activities that provide economic welfare to members of society.

Nominal and real GDP

Since it is physical quantities of output that yield people utility or happiness, it can be misleading to judge the economy's performance by looking at **nominal GDP**.

Table 15.6 presents a simple hypothetical example of a whole economy. Nominal GDP rises from £600 to £1470 between 1980 and 2013. If we take 1980 as the base year, we can measure **real GDP** in 2008 by valuing output quantities in 2013 using 1980 prices. Real GDP rises only from £600 to £860. This rise of 43 per cent in real GDP gives a truer picture of the extra quantity of goods made by the economy as a whole.

CASE 15.2

TAX EVASION, CRIME AND UNDER-REPORTING OF GDP

Gangster Al Capone, never charged with murder, was eventually convicted of tax evasion. Taxes are evaded by smugglers and drug dealers but also by gardeners, plumbers and everyone else doing things 'for cash'. Since GDP data are based on tax statistics, the 'hidden' economy is unreported and official statistics understate the true value of GDP. This undeclared economic activity is sometimes called the shadow or hidden economy.

Economists have various ways to estimate its size. One way is to count large-denomination banknotes in circulation. People with fistfuls of £50 notes are often engaged in tax evasion. When the euro was first launched, the decision to make the most valuable note €500 (more valuable than the \$100 bill) led to fierce discussion as to whether the euro would replace the dollar as the preferred currency of crooks. In 2010 British bank wholesalers withdrew it from circulation in the UK, but they still circulate in the Eurozone, where they acquired the nickname 'Bin Ladens' because everyone knew they existed but they were rarely seen.



top: © Aidart. bottom: © AcePixure

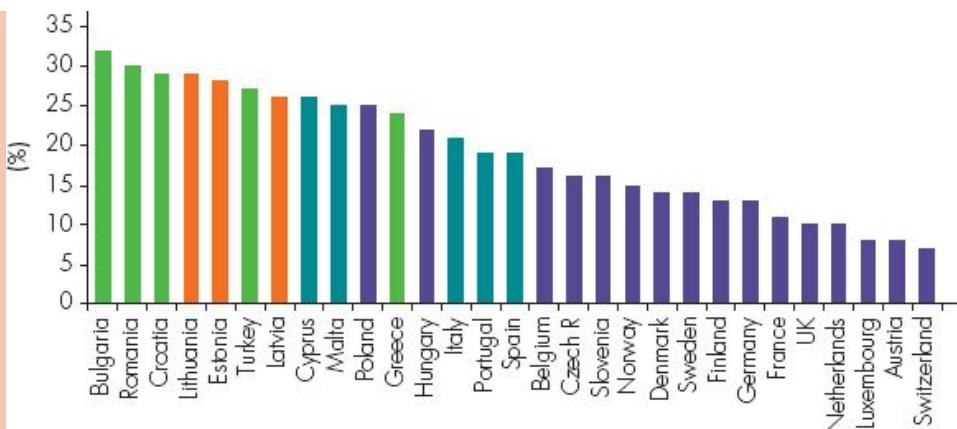
Another way is to guess people's income by studying what they spend. Maria Lacko exploited the stable relationship between household use of electricity and its main determinants – income and weather temperature – to estimate incomes from data on electricity consumption and temperature. She confirmed two popular views. The hidden economy is large both in former communist economies, where the new private sector is as yet unrecorded, and in several Mediterranean countries with a history of poor tax collection. She found that the size of the hidden economy might be around 20–30 per cent of reported GDP in the countries of Eastern Europe and the Mediterranean, but probably only 5–10 per cent of the size of GDP in the US and the UK. Measured properly, GDP would therefore be larger. Another way to estimate the hidden economy is to conduct surveys and offer people immunity if they tell the truth.

The broad thrust of these estimates has been confirmed in recent work by Friedrich Schneider, cited in 2012 by the EU Commission.

The hidden economy ranges from 7–8 per cent of GDP in law-abiding Switzerland and Austria to around 30 per cent in many of the Balkan countries (shown in green). Tax evasion and under-reporting of GDP is also prevalent in the Baltic states (shown in orange).

Why does a country have a large shadow economy? Partly, the answer lies in culture and history. Balkan states, ruled by the Ottoman Empire for 400 years, got used to evading central government; in contrast, Switzerland, with its recurring referenda consulting citizens on many policy issues, has much greater buy-in to government policy.

Economic arguments also play a role. When tax rates get too high, citizens become increasingly preoccupied with tax avoidance (legal) or tax evasion (illegal). How high tax rates have to be before inducing such antisocial behaviour is still a matter of controversy.



The hidden economy (% of measured GDP)

Source: F. Schneider, The size and development of the shadow economy from 2003–2012 (EU Commission, 2011).

Table 15.6 Nominal and real GDP

		1980	2013
Quantity	Apples	100	150
	Chickens	100	140
Price £s	Apples	2	4
	Chickens	4	6
Value in 2013 £s	Apples	200	600
	Chickens	400	840
	Nominal GDP	600	1440
Value in 1980 £s	Apples	200	300
	Chickens	400	560
	Real GDP	600	860

The GDP deflator

Chapter 2 introduced the consumer price index (CPI), an index of the average price of goods purchased by consumers. The most common measure of the inflation rate in the UK is the percentage rise in the CPI over its value a year earlier.

However, consumption expenditure is only one part of GDP, which also includes investment, government spending and net exports. To convert nominal GDP to real GDP, we need to use an index showing what is happening to the price of all goods. This index is called the **GDP deflator**.

The **GDP deflator** is the ratio of nominal GDP to real GDP expressed as an index.

Figure 15.4 plots data for the UK during 1998–2011, showing nominal GDP (in £ billions) in green. Looking at the green line alone, it appears that the recession of 2008/09 was short-lived, and that growth quickly resumed. The purple line plots UK

GDP in real terms, valuing it always at the prices ruling in 2009. The two curves therefore cross in 2009. Once we allow for the effect of inflation in boosting nominal GDP, we can see that real GDP grew much more slowly. The severity of the recession after 2008 then becomes much clearer.

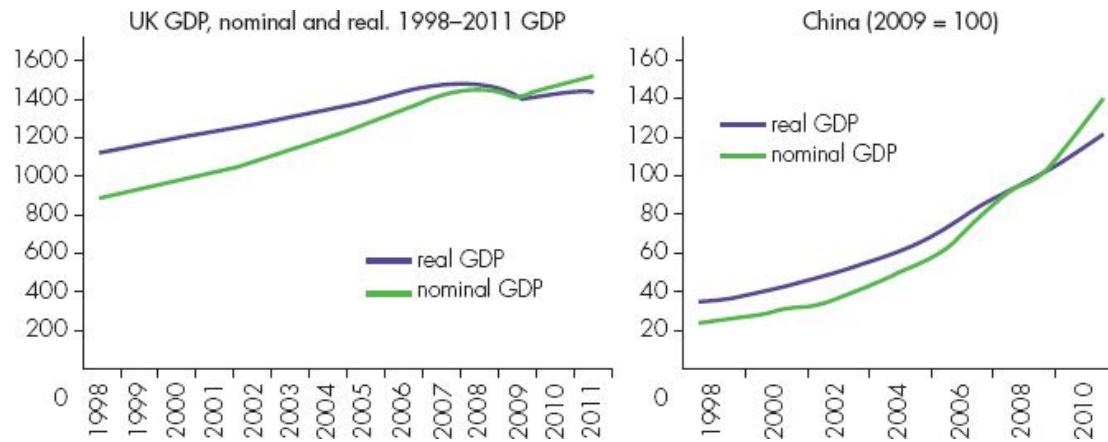


Figure 15.4 Nominal and real GDP

Source: IMF, World Economic Outlook, April 2013.

Figure 15.4 also shows indexes of nominal and real GDP in China over the same period. Because its real GDP growth was much stronger, the purple curve is much steeper than for the UK. Moreover, since China has much more inflation than the Eurozone or the UK, its nominal GDP growth is even greater still.

Per capita real GDP

Real GDP measures the real output of an economy. Its annual percentage increase tells us how fast an economy is growing. Table 15.7 shows average annual growth rates of real GDP during 2000–2010 for a range of countries. It also shows average annual population growth during the period. The difference between these two columns shows us the average annual growth of output per person.

Per capita real GDP is real GDP divided by total population. It is real GDP per head.

Chinese real GDP grew on average by 10.8 per cent a year. Because of its one child per family policy, its population hardly grew. Hence it achieved a stunning 10.2 per cent a year increase in **per capita** output, allowing its living standards to grow rapidly. Although Ethiopian GDP grew by 8.8 per cent annually, its population was also growing strongly. Output per person grew by 6.2 per cent. In the UK and Switzerland, annual GDP growth was less than 2 per cent, implying annual growth rates of around 1 per cent in per capita GDP. In the Palestine areas of the West Bank and Gaza, not only did real GDP fall over the period, rapid population growth of 3.2 per cent a year meant that per capita GDP was falling by 3.7 per cent a year during 2000–2010.

Table 15.7 Annual growth of real GDP and per capital real GDP, 2000–12 (% per annum)

	Real GDP	Population	Per capita real GDP
China	10.8	0.6	10.2
Ethiopia	8.8	2.6	6.2
Switzerland	1.9	0.8	1.1
UK	1.6	0.6	1.0
West Bank/Gaza	-0.9	3.2	-3.7

Whether we are interested in GDP or per capita GDP depends on the question we want to answer. For example, total GDP will give an indication of a country's economic size and power, whereas per capita GDP is more informative about productivity and likely living standards of the representative person. Of course, individuals do not always follow the average. The more the dispersion changes over time, the less reliable the average or per capita statistics are in thinking about what is happening to particular people.

A comprehensive measure of GDP

Because we use GDP to measure the output of the economy, the coverage of GDP should be as comprehensive as possible. In practice, we encounter two problems in including all production in GDP. First, some outputs, such as noise, pollution and congestion, are 'bads'. We should *subtract* them from GDP. This is a sensible suggestion but hard to implement. These nuisance goods are not traded through markets, so it is hard to quantify their output or decide how to value their cost to society.

Similarly, many valuable goods and services are excluded from GDP because they are not marketed and therefore hard to measure accurately. These activities include household chores, DIY activities and unreported jobs.

Deducting the value of nuisance outputs and adding the value of unreported and non-marketed incomes would make GNP a more accurate measure of the economy's production of goods and services. But there is another important adjustment to make before using GNP as the basis for national economic welfare. People enjoy not merely the consumption of goods and services but also leisure time.

ACTIVITY 15.1

SUSTAINABILITY, POLLUTION AND NEGATIVE GDP GROWTH

The first table below shows data on the successful industrialization of Asian economies. The Asian tigers comprise small countries in East Asia that abandoned protectionism in the 1960s and began half a century of rapid export-led industrialization.

Global manufacturing success of the Asian tigers

	Per capita real GDP annual growth (%), 1965–2012	Manufactured exports (% of total exports)	
		1965	2012
Indonesia	6	2	41
Malaysia	5	6	70
Singapore	7	34	74
South Korea	6.5	59	90
Thailand	6	4	75

Source: World Bank, World Development Report (various issues).

In the 1990s a second wave of countries embraced globalization, exports and significant deregulation. The most important of these – because of their size, their success has had a huge impact on the world economy – are now known as the BRICs (Brazil, Russia, India and China).

BRIC_s economic success

	GDP growth Annual average (%) 2000–12	Per capita GDP growth Annual average (%) 2000–12	Share of manufactures in total exports (%)	
			1980	2012
Brazil	3.5	3.5	39	39
Russia	5	5	–	17
India	8	3	59	67
China	10	10	48	94

Although China, and to a lesser extent India, industrialized in a manner similar to the Asian tigers, Russia is essentially an energy exporter and Brazil's growth (like Australia's) has largely reflected exports of raw materials to booming Asian manufacturers.

Whatever the differences between countries, the central point is that emerging markets are booming, industrializing quickly (even in raw material extraction) and developing large urban populations of sophisticated consumers.

From earlier industrial revolutions in Europe, the US and Japan, we know that this phase of economic development usually entails considerable pollution – the dark satanic mills of England, or the pall that used to hang over New York and Chicago. Environmental quality is a luxury of much greater concern to countries already secure in reasonable standards of material welfare. Rich and educated Europe now places much greater weight on the green agenda.

A study by the Asian Development Bank noted that Asian emissions of greenhouse gases will treble in the next 25 years. Asia will overtake Western economies as the world's biggest source of greenhouse gas pollutants. China is currently building a new coal-fired power station every two weeks, and India's microcar – the Tata Nano – will make motoring affordable to tens of millions of new drivers every year. As populations move from villages to the cities, demand for heat and power increases steadily.

Environmental degradation means that almost 40 per cent of Asia's population now lives in areas prone to drought and erosion. With the Asian population set to triple

in the next 20 years, and half these people living in cities, air pollution will reach new records. Nor is access to clean water much better.

If national accounts kept a proper account of environmental depreciation – a cost subtracted from gross output when measuring the true net output of an economy – many Asian countries would have less impressive growth records. We might even have to call them Asian snails instead of tigers. Their success in making consumer electronics is offset by extensive pollution and urban congestion. Ask anyone who recently visited Bangkok.

Questions

- a. How does depreciation of ordinary machinery and buildings enter calculations of GDP or national income?
- b. What national accounts measure properly reflects depreciation of physical capital?
- c. How are conventional estimates of depreciation made?
- d. What would be entailed in following the same procedures for environmental capital?
- e. How would environmental capital for the whole planet affect national accounts?

To check your answers to these questions, see page 679.

Suppose Leisurians value leisure more highly than Industrians. Industrians will therefore choose to work more than Leisurians, and therefore will also produce more goods and services. Industria has a higher measured GDP. It is silly to say this proves that Leisurians have lower welfare. By choosing to work less, they reveal that the extra leisure is worth at least as much as the extra goods and services they could have made by working more. Ideally, we should be measuring the output of leisure as well as of goods and services.

Because it is difficult and expensive to collect regular measurements on non-marketed and unreported goods and bads, and to make regular assessments of the implicit value of leisure, real GDP inevitably remains the commonest measure of economic activity. Far from ideal, it is the best measure available on a regular basis.

15.6 International comparisons

International agencies prefer to compare like with like, which means eliminating measures that are sensitive to large differences in national practices. No country invests much of its resources in collecting accurate data for depreciation. Hence international statistics focus on gross measures rather than net measures, since the latter would entail making allowances for depreciation, which vary from country to country.

Similarly, it is generally felt that GDP data are more reliable than GNP data, since the latter includes estimates of property income earned abroad. One reason why people hold assets abroad is to avoid declaring the income to national tax authorities. Assessing capital gains (which are really income) on foreign assets is also problematic. Hence, most

international comparisons are based on GDP. We have a more reliable idea of gross output than of gross income.

The new economics of happiness

Suppose we wanted to go beyond imperfect measures of material goods and services. How might we proceed? The new economics of happiness tries to ask people directly how happy they feel and then correlates this with various measures of their economic, social and psychological environment.

Simple economics might suggest that the average happiness of a citizen of a particular country should be correlated with per capita GDP in that country. There is quite a lot of empirical support for this correlation if we are talking about fairly poor countries – a higher standard of living makes a big difference. However, research by Richard Easterlin discovered that, beyond some threshold of material well-being, additional levels of material goods and services lead to surprisingly small additions to subjective happiness. Once basic needs have been met, people appear to derive more happiness from being *relatively* better off than other citizens. Nor is material standard of living the only relevant determinant – people report greater happiness if they are in a stable relationship with a partner, have a reasonable amount of freedom and political control over their lives, have plenty of leisure, enjoy good health and live in a society free from conflict and warfare.

The Satisfaction with Life index, based on survey data, allows the construction of a global happiness map. In Figure 15.5, purple denotes the happiest countries, and red the least-happy countries. There is obviously quite a strong correlation with per capita GDP. North America, Europe, Australia and New Zealand and Saudi Arabia all score highly, whereas Africa fares poorly. However, Russia is less happy than its per capita GDP would imply, whereas Mexicans appear as happy as Europeans, despite making and earning considerably less.

Amongst individual countries, Denmark, Switzerland and Austria topped the ranking, whereas Burundi, Zimbabwe and the Democratic Republic of the Congo came bottom. To keep track of the latest version of this index, use your favourite Internet search engine to look for the Economist Quality of Life index.

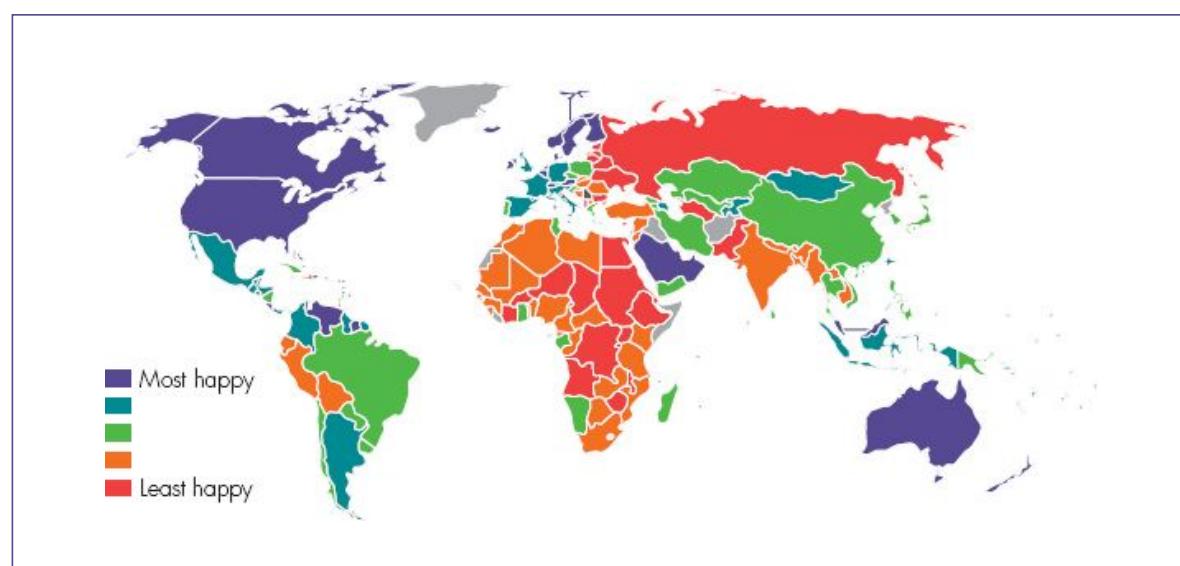


Figure 15.5 The Satisfaction with Life index

Source: A. White, 'A global projection of subjective well-being', Psychtalk 56 (2007): 1720.

Summary

- **Macroeconomics** examines the economy as a whole.
- Macroeconomics sacrifices individual detail to focus on the interaction of broad sectors of the economy. Households supply production inputs to firms that use them to make output. Firms pay factor incomes to households, who buy the output from firms. This is the **circular flow**.
- **Gross domestic product (GDP)** is the value of net output of the factors of production located in the domestic economy. It can be measured in three equivalent ways: value added in production, factor incomes including profits or final expenditure.
- **Leakages** from the circular flow are those parts of payment by firms to households that do not automatically return to firms as spending by households on the output of firms. Leakages are saving, taxes net of subsidies and imports. **Injections** are sources of revenue to firms that do not arise from household spending. Investment expenditure by firms, spending on goods and services by the government and exports are injections. By definition, total leakages equal total injections.
- **GDP at market prices** values domestic output at prices inclusive of indirect taxes.
GDP at basic prices measures domestic output at prices exclusive of indirect taxes.
Gross national product (GNP), also called gross national income (GNI), adjusts GDP for net property income from abroad.
- National income is net national product (NNP) at basic prices. NNP is GNP minus the **depreciation** of the capital stock during the period. In practice, many assessments of economic performance are based on GNP since it is hard to measure depreciation accurately.
- **Nominal GDP** measures output at current prices. **Real GDP** measures output at constant prices. It adjusts nominal GDP for changes in the **GDP deflator** as a result of inflation.
- **Per capita real GDP** divides real GDP by the population. It is a more reliable indicator of average output per person in an economy.
- Real GDP and per capita real GDP are crude measures of national and individual welfare. They ignore non-market activities, bads such as pollution, valuable activities such as work in the home, and production unreported by tax evaders. Nor do they measure the value of leisure.

- Because it is expensive, and sometimes impossible, to make regular and accurate measurements of all these activities, in practice GDP is the most widely used measure of national performance.
- By using data based on surveys of how happy people feel, it is possible to explore how reliable per capita GDP is as a measure of national happiness. There is a strong relationship, especially before an acceptable living standard has been reached, but happiness also depends on environmental factors (security, access, freedom) and other personal factors (health, success of relationships).

Review questions



EASY

1 Car firms buy raw materials (steel), intermediate goods (windscreens, tyres) and labour to make cars. Windscreen and tyre companies hire workers and also buy raw materials from other industries. What is the value added of the car industry (the three firms shown below)?

Producer of	Output	Intermediate goods used	Raw materials used	Labour input
Cars	1000	250	100	100
Windscreens	150		10	50
Tyres	100		10	30

2 GNP at market prices is £300 billion. Depreciation is £30 billion; indirect taxes £20 billion. (a) What is the national income? (b) Why does depreciation cause a discrepancy between GNP and national income? (c) Why do indirect taxes enter the calculation?

3 Which of the following statements is correct? (a) Increasing the staff of the police force in response to higher crime raises national income because government spending is higher. (b) Increasing the staff of the police force reduces national income because society wastes resources tackling crime. (c) There is no effect on national income because the benefit of more police is offset by the cost of more crime.

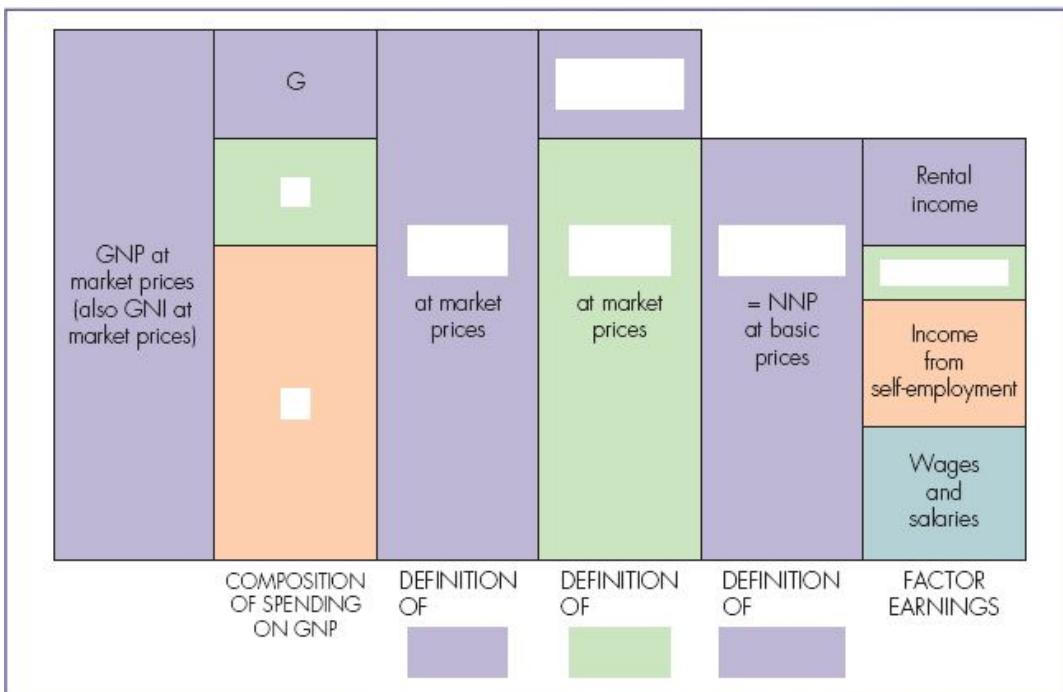
4 Given the data below: (a) What is 2010 GNP in 2009 prices? (b) What is the growth rate of real GNP from 2009 to 2010? (c) What is the inflation rate?

Year	Nominal GDP (£)	Income elasticity
2009	2000	100
2010	2400	100

5 **Common fallacies** Why are these statements wrong? (a) Unemployment benefit props up national income in years when employment is low. (b) A high per capita real GNP is always a good thing. (c) In 2010, *Crummy Movie* earned £1 billion more at the box office than *Gone with the Wind* earned 50 years ago. *Crummy Movie* is definitely a bigger box office success.

6 Suppose a country is unable to borrow from abroad and must always equate the value of its exports and imports. If the private sector is saving a lot more than it is investing, is the government in surplus or deficit? Why?

7 It is the year 2060. Nation states have been abolished and there is a world government whose spending is entirely financed by income tax. How does the diagram below have to be amended?



MEDIUM

8 Suppose the injections to the circular flow (investment I , government spending G and exports X) do not depend on the current level of national output Y . In contrast, suppose leakages increase as output increases. (a) Suppose saving $S = 0.1Y$. If total injections equal 10, what is the equilibrium level of output? (Assume a circular flow model without government and foreign sectors.) (b) Now suppose imports $Z = 0.4Y$ and taxes $T = 0.5Y$. If $G = £40$, is the government budget ($G - T$) in surplus or deficit?

9 GNP 5 £2000, C 5 £1700, G 5 £50 and NX 5 £40. (a) What is investment I ? (b) If exports are 350, what are imports? (c) If depreciation 5 £130, what is national income? (d) In this example, net exports are positive. Could they be negative?

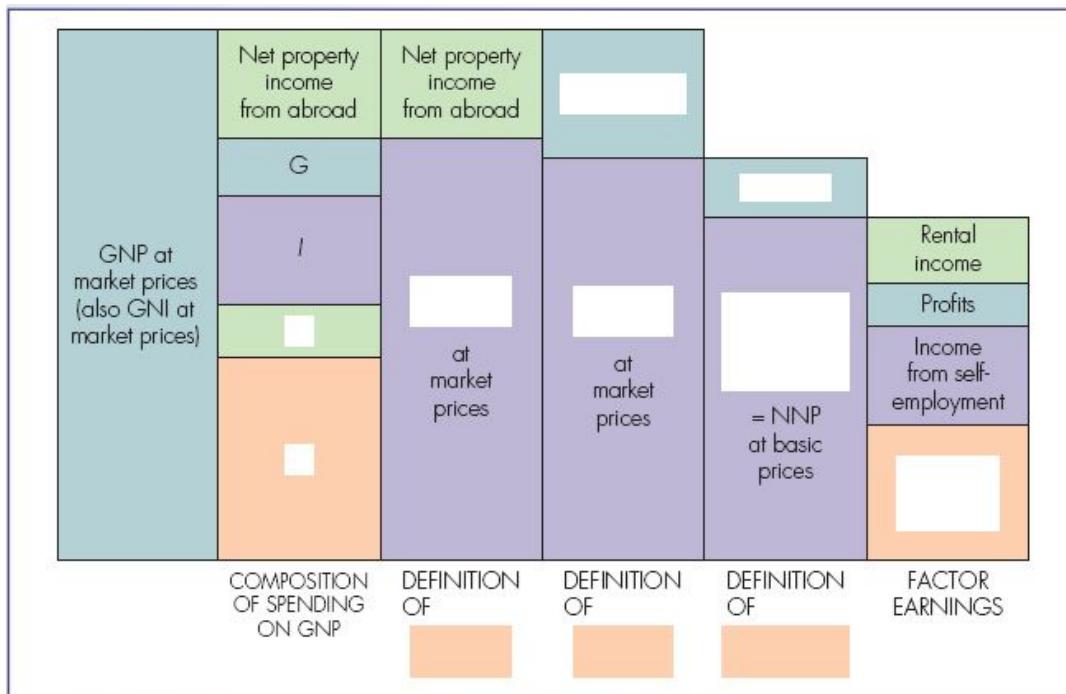
10 Should these be in a comprehensive measure of GNP? (a) Time spent by students in lectures. (b) The income of muggers. (c) The wage paid to traffic wardens. (d) Dropping litter.

11 The price of a new television has remained roughly constant for the last 30 years. What does this show?

HARD

12 Suppose the government publishes world economic accounts that estimate not merely depreciation of the stock of physical capital but also the stock of environmental capital. (a) Complete the composition of world income and output in the diagram given below. (b) If pollution and climate change are causing adverse effects, how does your new

diagram on world income and output differ from the diagram given below? (c) Suppose environmental depreciation was initially \$500 billion and the world government spends \$100 billion on pollution control with the consequence that environmental depreciation is now only \$300 billion. How must the diagram below be amended?



- |3 You are head of the Leisure Commission that has to recommend to the government how to include the value of leisure in GDP. How do you come up with an estimate?
- |4 **Essay question** ‘Economists are preoccupied with what they can measure. GDP is so misleading an indicator of welfare that it is almost pointless to gather statistics about it, either for international comparison across countries or to assess how well particular governments are doing.’ How useful is GDP? Could we easily have a better indicator?

1 It helps to draw parallels with microeconomics. The demand curve shows desired purchases at any price; the supply curve shows the desired sales at any price. In equilibrium, desired purchases equal desired sales. When the price is too high, there is excess supply and some desired sales are frustrated. But since every transaction has a buyer and a seller, actual purchases equal actual sales whether or not the market is in equilibrium.

CHAPTER 16

Output and aggregate demand

Learning outcomes

By the end of this chapter, you should be able to:

- 1 contrast actual output and potential output
- 2 explain why output is demand determined in the short run
- 3 define short-run equilibrium output
- 4 analyse consumption and investment demand
- 5 show how aggregate demand determines short-run equilibrium output
- 6 explain inflationary and deflationary gaps
- 7 define the marginal propensity to consume
- 8 calculate the multiplier
- 9 show how the marginal propensity to consume affects the multiplier
- 10 explain the paradox of thrift

GDP growth is seldom smooth. Using data from the very helpful International Monetary Fund (IMF) website, Figure 16.1 shows annual real GDP growth for the UK, Ireland, Sweden and the Netherlands since 1981. On average, Ireland grew most quickly, the other countries averaging about 2 per cent annual GDP growth. However, there were considerable fluctuations around this trend, even before the sharp collapse and recovery of output after the financial crash. Words used by economists to describe these fluctuations - recession, recovery, boom and slump – are part of everyday language.

Why does real GDP fluctuate? To construct a simple model, we ignore discrepancies between national income, real GNP and real GDP. We use income and output interchangeably. First, we distinguish between *actual* output and *potential* output.

Potential output tends to grow over time as the supply of inputs grows. Population growth adds to the labour force. Investment in education, training and new machinery adds to human and physical capital. Technical advances let given inputs produce more output. Together, these explain average annual growth of at least 2 per cent since 1981.

Potential Output is the economy's output when inputs are fully employed

We study the theory of long-run economic growth in potential output in Chapter 28. First, we focus on deviations of actual output from potential output in the short run. Since potential output changes slowly, we begin with a short-run analysis of an economy with a fixed potential output.

Potential output is not the maximum an economy can conceivably make. With a gun to our heads, we could all make more. Rather, it is the output when every market in the economy is in long-run equilibrium. Every worker wanting to work at the equilibrium wage can find a job, and every machine that can profitably be used at the equilibrium rental for capital is in use. Potential output includes an allowance for 'equilibrium unemployment'. Some people do not want to work at the equilibrium wage rate. Moreover, in a constantly changing economy, some people are temporarily between jobs. Today, UK potential output probably entails an unemployment rate of about 5 per cent, yet recent actual unemployment has exceeded 8 per cent.

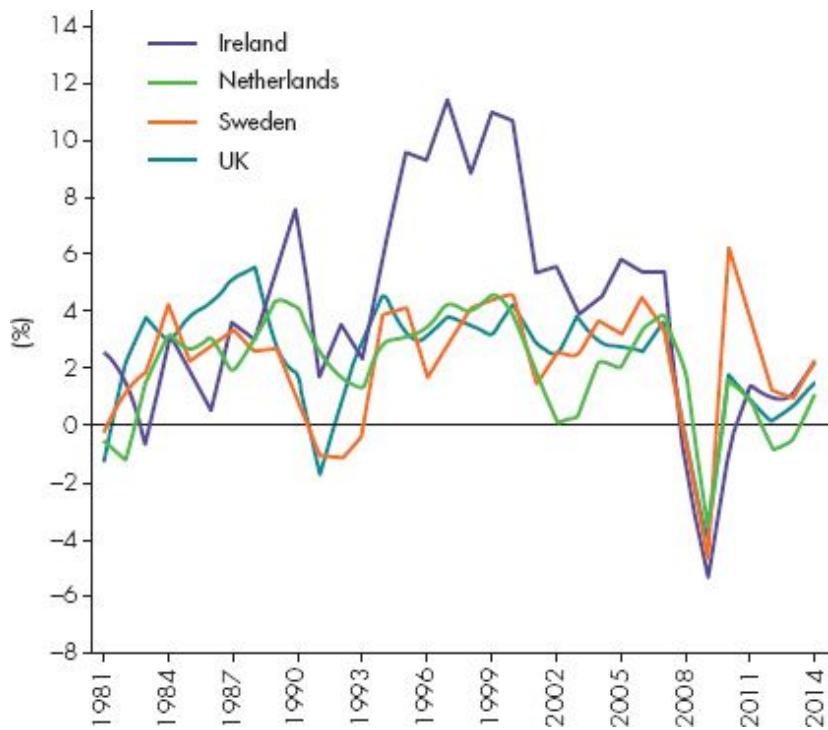


Figure 16.1 Annual real GDP growth, 1981-2014 (%)

Source: <http://www.imf.org/external/pubs/ft/weo/2013/01/weodata/index.aspx>

Suppose actual output falls below potential output. Workers are unemployed and firms have idle machines or spare capacity. A key issue in macroeconomics is how quickly output returns to potential output. In microeconomics, studying one market in isolation, we assumed excess supply would quickly bid the price down, eliminating excess supply to restore equilibrium. In macroeconomics, this cannot be taken for granted. Disturbances in one part of the economy induce changes elsewhere that may feed back again, exacerbating the original disturbance.

We cannot examine this issue by *assuming* that the economy is always at potential output, for then a problem could never arise. We must build a model in which departures from potential output are possible, examine the market forces then set in motion and decide how successfully market forces restore output to potential output. Because we want initially to focus on the possibility of additional unemployment, we start by considering a world in which there might be idle capacity and too little demand.

Trade is voluntary: actual exchange is always the smaller of supply and demand. Output is **demand-determined** if there is excess supply. Wages and prices have yet to adjust to restore long-run equilibrium. Until then, output depends only on aggregate demand.

Conversely, if excess demand exists, as under rationing in the former USSR, output is **supply-determined**.

Thus our initial model has two crucial properties. First, all prices and wages are fixed at a given level. Second, at these prices and wage levels, there are workers without a job who would like to work, and firms with spare capacity they could profitably use. The economy has spare resources. It is then unnecessary to analyse the supply side of the economy in detail. Any rise in demand is happily met by firms and workers until potential output is reached.

Below potential output, firms happily supply whatever output is demanded. Total output is **demand-determined**. (Conversely, if excess demand exists, as under rationing in the former USSR, output is **supply-determined**.)

Later, we relax the assumption that prices and wages are fixed. Not only do we want to study inflation, we also want to examine how quickly market forces, acting through changes in prices and wages, can eliminate unemployment and spare capacity. But first we must learn to walk. We defer analysis of price and wage adjustment until Chapter 21.

CASE 16.1

A BRIEF HISTORY OF MACROECONOMICS

Two of the most fundamental questions in macroeconomics are whether, and if so how quickly, we can rely on markets to restore actual output to potential output. If this happens reliably and quickly, output is usually near potential output; most economic analysis and economic policy should then focus on how potential output increases over time. However, if actual output can deviate from potential output for long periods and by large amounts, we must focus much more on what determines actual output.

Classical economists of the eighteenth and nineteenth centuries were market optimists, believing that market forces restore actual output to potential output quite quickly. The supply of potential output then became the principal focus of study. French economist Jean-Baptiste Say (1767–1832) gave his name to 'Say's Law', which states that supply creates its own demand. If goods and services are being produced, and the income thereby derived is passed on to the owners of the factors of production, the latter then have the spending power to purchase the original output. Nineteenth-century UK economists in the classical tradition include James Mill, his son John Stuart Mill, and David Ricardo. Classical economists still had some explaining to do – for example, why was there a Great Depression during 1873–79 in many leading economies? Had potential output really fallen by that much that quickly?

Nevertheless, classical economics largely persisted until the 1920s. After the First World War, European economies were struggling to regain pre-war prosperity. At the time, they all belonged to the gold standard, which in effect was a fixed exchange rate system. Within this system, countries such as the UK had become seriously uncompetitive as a result of wartime inflation. The 1926 General Strike in the UK was the result of austerity designed to push wages down to restore UK competitiveness, a close analogy of modern Greek misery as it tries to reduce its budget deficit and increase its competitiveness in order to survive within the Eurozone.

Less ravaged by the war and unburdened by wartime debts, the US boomed in the late 1920s, in ways very similar to the run-up to the financial crash of 2008: an asset price bubble in the stock market and housing market, excessive borrowing and injudicious lending. The crash of 1929 tipped the locomotive of the world economy into its own protracted recession. The 1930s became known as the Great Depression. The explanation offered by classical economics became less and less convincing.

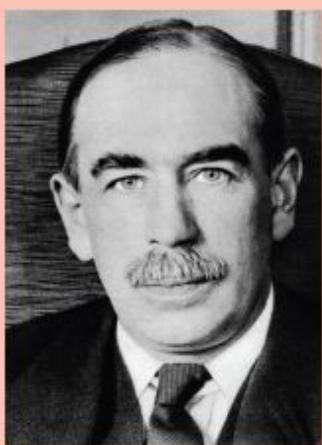
The intellectual revolution was led by UK economist John Maynard Keynes, culminating in his *General Theory of Employment, Interest and Money* (1936). Keynes used the model to explain high unemployment and low output in the Great Depression of the 1930s. Discarding classical assumptions, Keynes argued that market forces – the adjustment of wages and prices in response to booms and

slumps – were too slow and too ineffective to prevent serious and prolonged departures of demand and output from potential output.

Many young economists soon became *Keynesians*, advocating government intervention to manage demand and output, keeping them close to potential output. This approach dominated economic policy and intellectual thinking in the 1950s and 1960s, when governments were proud of their expertise in keeping output growth smooth and unemployment low.

However, the magic began to wear off by the late 1960s. Inflation was steadily rising and the normal level of unemployment was creeping up, even when governments boosted demand. Indeed, the pursuit of 'artificially low' unemployment, and failure to recognize that the underlying sustainable level of unemployment had increased, was a principal cause of the rising inflation.

Monetarists, led by Chicago economist Professor Milton Friedman, argued that we could at least take care of inflation, by pursuing a prudent rate of money growth. In essence, this meant abandoning the attempt to micromanage demand and output and relying to a greater degree on market forces again to restore output slowly to potential output. Friedman famously argued that the lags in macroeconomic policy were 'long and variable'. Even well-meaning interventions took an unpredictable time to take effect, by which time the economy might have sorted itself through market forces. Active macroeconomic policy risked blowing the economy off course just as it was recovering unaided.



© ClassicStock | Alamy

The monetarist revival also reminded everyone that incentives and 'supply-side policies' matter – whatever the path of actual output, we

still want a good evolution of potential output and the corresponding level of normal or 'equilibrium' unemployment.

Following the sharp rise in oil prices achieved by OPEC, in the 1970s output stagnated, unemployment rose and inflation was initially obdurately high. Some economists discarded Keynesian economics completely. Not only did they deny the effectiveness of government policy to stabilize output, they argued that stabilizing output might not even be desirable. This prompted a fightback by *New Keynesians*, who believe that the central messages of Keynes, right all along, can be understood better by using modern microeconomics to explain the market failures that justify Keynesian intervention to assist the effective operation of markets.

After the mid-1990s, there appeared to be considerable convergence in ideas about macroeconomics. Central banks were made independent from government interference and asked to stabilize national economies, particularly their inflation rates. Central banks appeared to combine responsible underlying monetary growth with active fine-tuning of interest rates to keep output close to potential output. This appeared to confirm (a) that, properly managed and without political interference, demand management could succeed; and (b) that active demand management could be combined with long-run inflation control. For nearly two decades, this appeared to work well.

Perhaps it all worked too well. In stabilizing economies, central banks eliminated a lot of risk for the private sector. In a climate of safety, the private sector felt able to take huge 'risks' precisely because it did not think they were risky. Households borrowed too much and speculated on property, banks borrowed too much and invested in more and more dubious assets in pursuit of ever-higher returns. The central banks were so focused on macroeconomic stability of inflation and GDP that nobody blew the whistle on increasingly dangerous private sector behaviour. The result was the crash of 2008.

Clawing our way out of this large hole is proving extremely difficult and slow. As asset prices collapsed, banks and their borrowers approached bankruptcy. Sensibly, governments stepped in to prevent the crisis spiralling, but at the cost of transferring the debts to the governments themselves. This has severely constrained future policy.

With everyone fearful, few now wish to spend. Without their demand, firms are reluctant to supply.

Mastering modern macroeconomics will take us to great heights and shed light on these issues. But we need to climb slowly, and to begin at the foothills.

Chapter 15 introduced the circular flow of income and payments between households and firms. Households buy the output of firms. Firms' revenue is ultimately returned to households. We now build a simple model of this interaction of households and firms. Then, in Chapter 17, we continue the analysis after adding the government and the foreign sector.

16.1 Components of aggregate demand

Without a government or a foreign sector, there are two sources of demand: consumption demand by households and investment demand by firms. Aggregate demand AD is the sum of consumption demand C and investment demand I . Consumption demand and investment demand are chosen by different economic groups and depend on different things.

Consumption demand

I Households buy goods and services from cars to cinema tickets. These consumption purchases account for most of **personal disposable income**.

Personal disposable income is the income households receive from firms, plus transfers received from government, minus taxes paid to government. It is the net income households can spend or save.

With no government, disposable income is simply the income received from firms. Given its disposable income, each household plans how much to spend and to save. Deciding one, decides the other. One family may save to buy a bigger house; another may spend more than its income, or 'dissave', taking the round-the-world trip it always wanted.

Many things affect consumption and saving decisions. We examine these in detail in Chapter 20. To get started, one simplification takes us a long

way. We assume that, in the aggregate, households' consumption demand rises with aggregate personal disposable income.

Figure 16.2 shows real consumption and real GDP, both corrected for inflation, during 1948–2011. The figure confirms that the basic relation between income and consumption is strong and stable over time. Each extra £1 of GDP was accompanied by around £0.7 extra consumption.

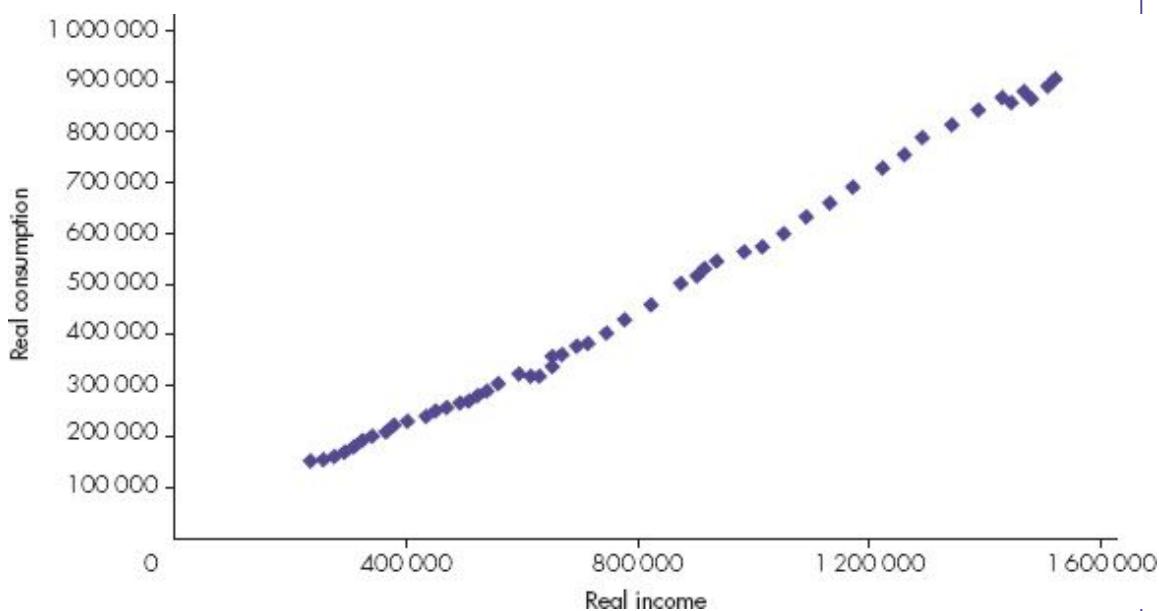


Figure 16.2 UK consumption and GDP, 1948–2011 (£m, 2009 prices)

Source: www.ons.gov.uk.

The consumption function

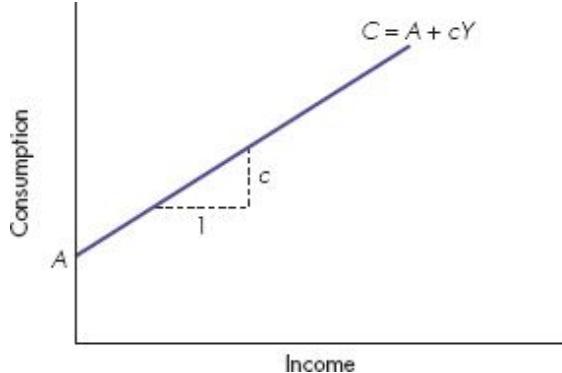
This positive relation between income and consumption demand is shown in Figure 16.3 and is called the **consumption function**.

The **consumption function** shows aggregate consumption demand at each level of income.

The consumption function tells us how to go from income Y to consumption demand C . If A is a positive constant, and c is a positive fraction between 0 and 1, then

$$C = A + cY$$

Our bare-bones model has no government, no transfer payments and no taxes. The consumption function is a straight line. A straight line is completely described by its intercept – the height at which it crosses the vertical axis – and its slope – the amount it rises for each unit we move horizontally to the right.



The consumption function shows aggregate consumption demand at each aggregate income. With zero income, autonomous consumption is A . The marginal propensity to consume c is the slope of the line, the fraction of each extra pound that households wish to spend. The remaining $(1 - c)$ they wish to save.

Figure 16.3 The consumption function

The intercept is A . We call this *autonomous* consumption demand. Since our basic model is that income determines consumption demand, autonomous means those determinants of consumption demand other than income. Households wish to consume A even if income Y is zero.¹

We might think of autonomous demand as reflecting the minimum consumption needed for survival, but the reality is more subtle. Autonomous consumption demand can change in response to changes in other economic variables, for example expectations about *future* incomes. Students expecting prosperous futures may spend more than is justified by their current incomes alone! The key point for our current analysis is that autonomous consumption demand does not depend upon current income.

The slope of the consumption function is the **marginal propensity to consume**.

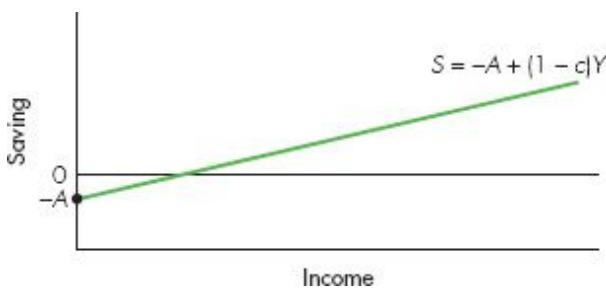
Marginal propensity to consume is the fraction of each extra pound of disposable income that households wish to consume.

Different people may exhibit different marginal propensities to consume. Poor people, with many unmet needs, are likely to spend immediately any extra income that they receive. For them, c is close to 1. Billionaires may already be consuming everything they could possibly want. For them, any extra income is largely unspent: c is close to 0. In macroeconomics, we are interested in aggregate behaviour, so the MPC will be less than 1 but well above 0.

In Figure 16.3 the marginal propensity to consume MPC is c . If income rises by £1, desired consumption rises by £ c . In Figure 16.2, c was around 0.6. Different people may behave differently. When poor people get extra income, they tend to consume almost all the extra. When a billionaire gets extra income, his consumption demand may be unchanged. The aggregate consumption function reflects average behaviour for the population as a whole.

Saving is income not consumed. When income Y is zero, saving is $-A$. Households are dissaving, or running down their assets. Since a fraction c of each pound of extra income is consumed, a fraction $(1 - c)$ of each extra pound of income is saved. The marginal propensity to save MPS is $(1 - c)$. Since an extra pound of income leads either to extra desired consumption or to extra desired saving, $MPC + MPS = 1$. The three-line identity symbol means that MPC plus MPS necessarily equals 1 as a matter of definition. Figure 16.4 shows the **saving function** corresponding to the consumption function in Figure 16.3.

The **saving function** shows desired saving at each income level.



The saving function shows desired saving at each income level. Since all income is saved or is spent on consumption, the saving function can be derived from the consumption function or vice versa.

Figure 16.4 The saving function

Since saving is defined as the part of income unspent, $Y \equiv C + S$, the three-line symbol promising us that this is always true as a matter of definition. We can replace C by the consumption function to deduce the relation between desired saving and income, the saving function shown in Figure 16.4.

At an income of zero, autonomous consumption is A . With income Y at zero, desired saving must therefore be $-A$. Since each unit increase in income leads to an extra c of desired consumption, it must also lead to an extra $(1 - c)$ of desired saving. Whatever is not consumed must be saved. Hence the saving function is as shown in Figure 16.4. Planned saving is the part of income not planned to be spent on consumption.

Sometimes we will refer to the *saving rate* S/Y , which is simply the fraction of income saved.

Investment spending

Income is the key determinant of household consumption or spending plans as described by the consumption function. What about the factors determining the investment decision by firms; that is, their planned spending on new capital goods in the form of factories, machinery and buildings?

Firms' **investment demand** depends chiefly on firms' current guesses about how fast the demand for their output will increase. Sometimes output is high and rising, sometimes it is high and falling. Since there is no close connection between the current *level* of income and firms' guesses about how the demand for their output is going to *change*, we make the simple assumption that investment demand is autonomous. Desired investment I is constant, independent of current output and income. In Chapter 20 we discuss investment demand in more detail.

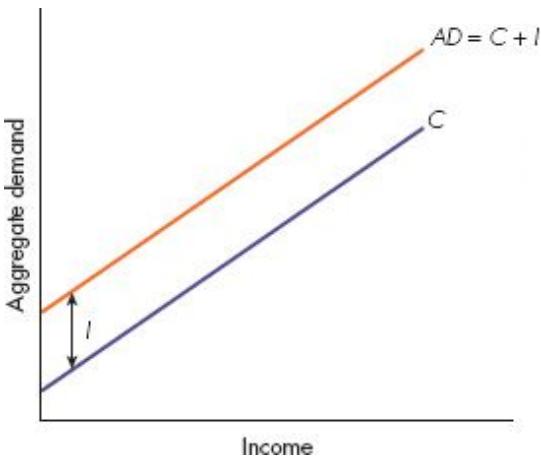
Investment demand is firms' desired or planned additions to physical capital (factories and machines) and to inventories.

16.2 Aggregate demand

In our simple model, **aggregate demand** is simply households' consumption demand C plus firms' investment demand I .

Aggregate demand is the amount firms and households plan to spend at each level of income.

Figure 16.5 shows the aggregate demand schedule. To the previous consumption function it adds a constant amount I for desired investment. Each extra unit of income adds c to consumption demand but nothing to investment demand: aggregate demand rises by c . The AD schedule is parallel to the consumption function. The slope of both is the marginal propensity to consume.



Aggregate demand is what households plan to spend on consumption and firms plan to spend on investment. Since we assume investment demand is constant, consumption is the only part of aggregate demand that increases with income. Vertically adding the constant investment demand to the consumption function C gives the aggregate demand schedule AD .

Figure 16.5 Aggregate demand

CONCEPT 16.1

OPPORTUNITY COSTS AND CHOICES: AN EXAMPLE

A model is like a sausage machine. Our economic theory is the design of how the machine works, which is deduced from our assumptions about how people behave. Even once it has been built, a sausage machine still needs inputs - raw meat, breadcrumbs, spices and the butcher's secret ingredients – in order to deliver an output of sausages. In the same way, our economic models require inputs of

some economic variables in order then to deliver implications for how other variables will behave.

Exogenous variables are those fed into the model as inputs.

Endogenous variables are those which the model then delivers as outputs, conditional on the values of the exogenous inputs.

In our simple model of aggregate demand, the levels of both investment and autonomous consumption are **exogenous**, or given from outside the model. Conditional on these inputs to the model, the model then determines the **endogenous variables**: consumption demand, and thus total aggregate demand.

16.3 Equilibrium output

Wages and prices are *fixed*, and output is demand determined. If aggregate demand falls below potential output, firms cannot sell as much as they would like. There is *involuntary* excess capacity. Workers cannot work as much as they would like. There is *involuntary* unemployment.

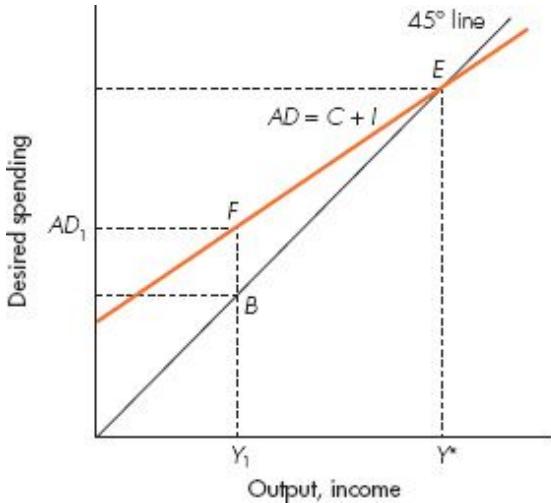
To define **short-run equilibrium** we cannot use the definition used in microeconomics, the output at which both suppliers and demanders are happy with the quantity bought and sold. We wish to study a situation in which firms and workers would like to supply more goods and more labour. Suppliers are frustrated. At least we can require that demanders are happy.

When prices and wages are fixed, at **short-run equilibrium output** aggregate demand or planned spending equals the output actually produced

Thus, spending plans are not frustrated by a shortage of goods. Nor do firms make more output than they can sell. In short-run equilibrium, actual output equals the output demanded by households as consumption and by firms as investment.

Figure 16.6 shows income on the horizontal axis and planned spending on the vertical axis. It also includes the 45-degree line, along which quantities

on the horizontal and vertical axes are equal.



The 45° line reflects any value on the horizontal axis on to the same value on the vertical axis. The point E, at which the AD schedule crosses the 45° line, is the only point at which aggregate demand AD is equal to income. Hence E is the equilibrium point at which planned spending equals actual output and actual income.

Figure 16.6 The 45° diagram and short-run equilibrium output

We draw in orange the aggregate demand schedule from Figure 16.5. This crosses the 45-degree line at E. On the 45-degree line, the value of output (and income) on the horizontal axis equals the value of spending on the vertical axis. Since E is the *only* point on the AD schedule also on the 45-degree line, it is the only point at which output and desired spending are equal.

Hence Figure 16.6 shows equilibrium output at E. Firms produce Y^* . That output is equal to income. At an income Y^* , the AD schedule tells us the demand for goods is also Y^* . At E, planned spending is exactly equal to the output produced.

At any other output, output is not equal to aggregate demand. Suppose output and income are only Y_1 . Aggregate demand exceeds actual output. There is excess demand. Spending plans cannot be realized at this output level.

Figure 16.6 shows that, for all outputs below the equilibrium output Y^* , aggregate demand AD exceeds income and output. The AD schedule lies *above* the 45-degree line along which spending and output are equal.

Conversely, at all outputs above the equilibrium output Y^* , aggregate demand is less than income and output.

ACTIVITY 16.1

THE AD SCHEDULE: MOVING ALONG IT OR SHIFTING IT?

The aggregate demand *AD* schedule is a straight line whose position depends on its intercept and its slope. The intercept, the height of the schedule when income is zero, reflects autonomous demand: the autonomous part of consumption demand and all of investment demand. The slope of the schedule is the *MPC*. Changes in income induce movements *along* a given *AD* schedule.

Autonomous demand is influenced by many things that we study in Chapter 20. It is not fixed for ever. But it *is* independent of income. The *AD* schedule separates out the change in demand directly induced by changes in income. All other sources of changes in aggregate demand are shown as *shifts* in the *AD* schedule. If firms get more optimistic about future demand and invest more, autonomous demand rises. The new *AD* schedule is parallel to, but higher than, the old *AD* schedule. The entire schedule has shifted upwards.

Mathematically, since aggregate demand is $AD = (A + I) + cY$, the level of autonomous demand ($A + I$) shows the height of the aggregate demand schedule when income is zero. Hence, changes in the autonomous components of consumption or investment change the height of the schedule, leading to a parallel shift. The slope of the schedule – how aggregate demand changes as income changes – is the marginal propensity to consume c . Movements in income Y lead to movements along a schedule of given height (since neither A nor I has changed). Finally, a change in the marginal propensity to consume c changes the slope of the *AD* schedule, causing it to rotate around the point on the vertical axis at which income is zero.

Questions

In each case, decide whether the *AD* schedule is shifting or whether the economy is moving along a given *AD* schedule:

- (a) As the Eurozone crisis continued, there was a wave of pessimism among UK consumers, who decided to save a lot more in order

to pay off the debts they had previously accumulated.

- (b) UK consumer spending has risen because households are having a good year and enjoying high incomes.
- (c) The 2012 Olympic Games in London caused an investment boom in UK construction.

To check your answers to these questions, go to page 679.

Adjustment towards equilibrium

In Figure 16.6, suppose that the economy begins with an output of Y_1 , below equilibrium output Y^* . Aggregate demand AD_1 exceeds output Y_1 . If firms have inventories from the past, they can sell more than they have produced by running down stocks for a while. Note that this destocking is *unplanned*; planned changes of stocks are already included in the total investment demand I .

If firms cannot meet aggregate demand by unplanned destocking, they must turn away customers. Either response - unplanned destocking or turning away customers - is a signal to firms to raise output above Y_1 . Similarly, at *any* output below Y^* , aggregate demand exceeds output and firms get signals to raise output.

Conversely, if output is initially above its equilibrium level, Figure 16.6 shows that output will then exceed aggregate demand. Firms cannot sell all their output, make *unplanned* additions to inventories and respond by cutting output.

Hence, when output is below its equilibrium level, firms raise output. When output is above its equilibrium level, firms reduce output. At the equilibrium output Y^* , firms sell all their output and make no unplanned changes to their stocks. There is no incentive to change output.

In this example, short-run equilibrium output is Y^* . Firms sell all the goods they produce, and households and firms buy all the goods they want. But nothing guarantees Y^* is the level of potential output.

The economy can end up at a short-run equilibrium output below potential output, with no forces then present to move output to potential output. At the given level of prices and wages, a lack of aggregate demand will prevent expansion of output above its short-run equilibrium level.

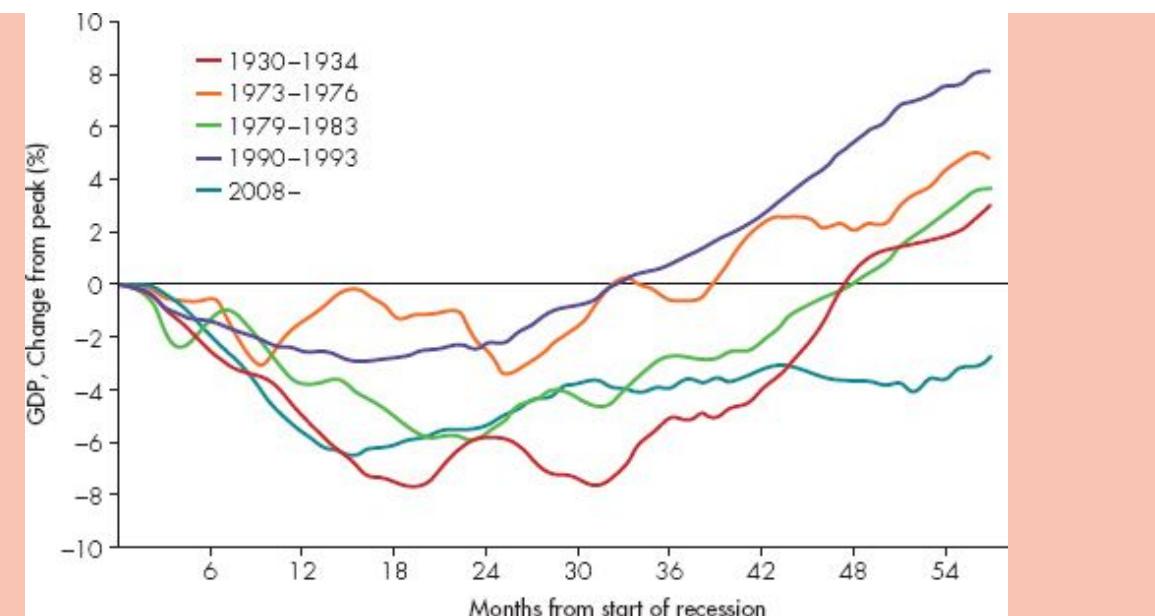
CASE 16.2

THE LITTLE DEPRESSION AND THE GREAT DEPRESSION

Economic historians teach their students about the Great Depression of the 1870s. Nowadays, they often call it the Long Depression, to distinguish it from the depression of the 1930s, which we have all come to call the Great Depression. However, at least in Europe, we may have to start calling the 1930s the Little Depression to distinguish it from the Great Depression that began in 2008.

We are all familiar with the terms of the business cycle – boom and bust, growth and recession, forwards and backwards. Does depression just mean recession or does it mean something else? A recession denotes two consecutive quarters of negative GDP growth; a depression is a protracted period in which output is significantly below potential output.

Thus, a depression is initially a deep recession, from which output then fails to recover in the normal way. The following chart is worth a thousand words. Constructed by the London-based National Institute for Economic and Social Research, it compares five UK recessions since 1930, plotting along the horizontal axis the number of elapsed months from the previous output peak, and on the vertical axis the cumulative change in GDP since the original start of the recession. Thus, a country exits recession once the curve's slope turns up (output is no longer falling), but output does not return to its previous level until the curve climbs back to 0 per cent.



The evolution of output during five UK recessions

Source: www.niesr.ac.uk

Since all the curves begin at 0 per cent on the left-hand side, we discuss the curves according to their height at the right-hand side. The purple curve shows the recession of 1990–93 when the UK pegged its exchange rate and interest rate to European countries at a time that just happened to include German unification. German interest rates were sky high to prevent inflation exploding in Germany, and the UK could not live with these high interest rates. On 'Black Wednesday' in September 1992, the UK abandoned the experiment, allowed its exchange rate to fall and slashed interest rates – economic recovery soon followed. Although output growth was quickly restored, it was not until three years (36 months) after the start of the crisis that output climbed back to its original level.

The orange curve depicts the recession after 1974 caused by a huge jump in oil prices after OPEC first became effective. Again, output was growing steadily within 24 months of the onset of the crisis, and had got back to its original level within around 36 months. Similarly, the green curve shows the sharp recession, but quick recovery, after Mrs Thatcher took office in 1979 and squeezed the economy to defeat inflation. Again, output was growing within 24 months, though it took 48 months to get back to the initial level of output.

The purple, orange and green curves all depict fairly conventional recession and recovery. The red curve depicts the Little Depression of the 1930s. It displayed the deepest fall in output, twice down

nearly 8 per cent from its initial level, and was clearly a 'double dip' since the initial recovery after 18 months was not sustained, leading to a new dip after 24 months. Even so, recovery took hold after 32 months and never looked back thereafter.

Finally, we come to the Great Depression since 2008 shown by the turquoise curve. Despite the enormity of the adverse shock – and the fact that it hit all major economies simultaneously – UK output initially fell by less than in the 1930s, in part because governments, led by the US and the UK, engaged in emergency expansionary measures to stave off the worst of the crisis. However, 48 months into the crisis, the turquoise curve is the *only* one still below its original level at the start, and as yet with no clear signs of getting back to its initial position. A third, or triple, dip is still a real possibility.

Thus, if we scale depressions by the length of time for which output is depressed, the most recent experience is not just the worst 'since the 1930s', it is already much worse 'than the 1930s'.

CASE 16.3

HOW DID THE CRASH AFFECT INVESTMENT?

Our simple model of aggregate demand assumes that output is the principal driver of consumption demand but does not directly affect investment demand. This does not mean that investment demand is always constant; merely that it is not well explained by changes in income. In later chapters we return to the question of what does affect investment demand.

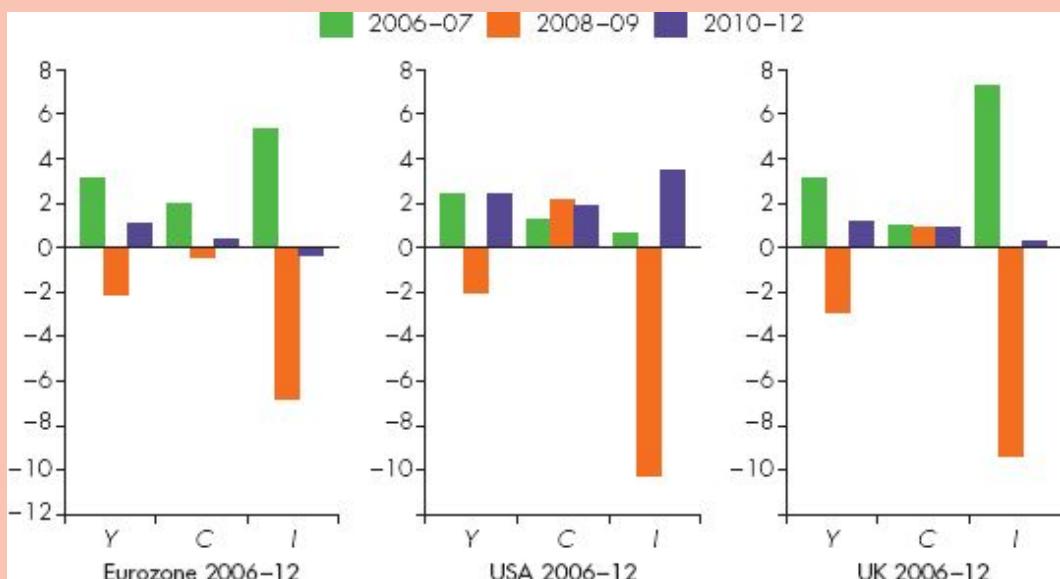
Even at this early stage, it is a good idea to check our theory is proceeding along the right lines. How did consumption and investment respond during the crash of 2009? The figure below shows data during 2006–12. It shows annual percentage changes in output, consumer spending and investment, and compares the Eurozone, the US and the UK.

In each figure below, the first three bars show changes in real output, the middle three show changes in real consumption, and the final three show changes in real investment. In the Eurozone, the US and

the UK, output Y was growing by at least 2 per cent a year during 2006–07 prior to the crash, then fell to around -2 per cent a year during 2008–09, before resuming slow growth after 2010. As we would expect from our discussion of the consumption function, these changes in output induced similar, but smaller, changes in consumption demand.

However, the behaviour of investment was very different. First, it certainly was not constant – it changed by much more than consumption demand. Investment demand is more volatile. Second, when the crash came, firms slashed investment spending by large amounts, and not merely because current income and output had fallen. Firms foresaw that the next few years were going to be tough, and that there would be little need for additional capital goods since production levels were not going to grow rapidly.

Thus, the figures support the basic idea of the consumption function – a close relation between current output and current consumption demand – but denies any similar relationship will work between output and investment demand. For now, we assume that investment is part of autonomous demand, independent of the level of output. Other things equal, investment is constant. But other things are not always equal, and actual investment can be highly volatile. We return to models of investment demand in later chapters.



Changes in output, consumption and investment, 2006–12 (annual averages, inflation adjusted)

Source: OECD, Economic Outlook, 2012.

16.4 Planned saving equals planned investment

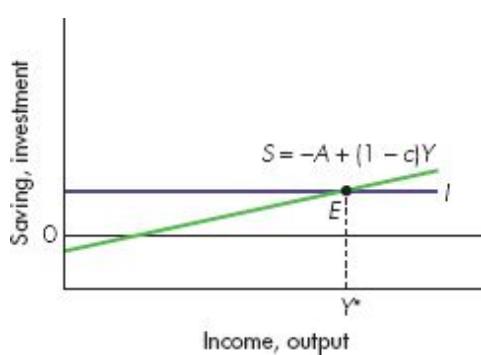
Equilibrium output equals the demand from investment and consumption: $Y = C + I$. This is not a definition, but holds only when output and income are at the right level to achieve equilibrium output. However, planned saving S is defined as the part of income Y not devoted to planned consumption C . Thus, $Y = C + S$. Together with $Y = C + I$, this implies that in equilibrium, but only in equilibrium, planned investment equals planned saving

$$I = S$$

In modern economies, firms make investment decisions, and the managers of these firms are not the same decision units as the households making saving and consumption plans. But household plans depend on their income. Since planned saving depends on income but planned investment does not, equilibrium income adjusts to make households plan to save as much as firms are planning to invest. Figure 16.7 illustrates.

Planned investment I is autonomous, and so a horizontal line since its magnitude is independent of the level of income. Planned saving increases with income and output, since the marginal propensity to save exceeds zero. Hence, equilibrium output must be Y^* , the only output at which planned investment equals planned saving.

Suppose investment demand is 30 and the saving function is $S = -10 + 0.4Y$. Hence, equilibrium output Y is 100. At this Y , planned saving is $[-10 + 40] = 30$. Hence 30 is both planned saving and planned investment.



At equilibrium output Y^* , planned investment I equals planned saving $S = -A + (1 - c)Y$.
Hence equilibrium output $Y^* = [A + I]/[1 - c]$.

Figure 16.7 At equilibrium output, planned I equals planned S

If the saving function is $S = -10 + 0.4Y$, the consumption function must be $C = 10 + 0.6 Y$. At an income of 100, consumption demand is 70. Add on 30 for investment demand, and aggregate demand is 100, just equal to equilibrium output of 100.

If income exceeds 100, households want to save more than firms want to invest. But saving is the part of income not consumed. Households are not planning enough consumption, together with firms' investment plans, to purchase all the output produced. Unplanned inventories pile up and firms cut output. Lower output and income reduces planned saving, which depends on income. When output falls back to 100, planned investment again equals planned saving.

Conversely, when output is below its equilibrium level, planned investment exceeds planned saving. Together, planned consumption and planned investment exceed actual output. Firms make unplanned inventory reductions and raise output until it reverts to its equilibrium level of 100.

MATHS 16.1

AUTONOMOUS DEMAND AND EQUILIBRIUM OUTPUT

In equilibrium, output equals aggregate demand, Hence

$$Y = AD = C + I = [A + cY] + I = [A + I] + cY$$

Hence, in equilibrium

$$Y^* = [A + I]/(1 - c) \quad (1)$$

Notice that this implies that a unit increase in either A or I then leads to an increase of $[1/(1 - c)]$ in equilibrium output Y^* . Since c is a positive fraction, $[1/(1 - c)]$ is greater than 1. So a unit increase in either autonomous consumption demand or investment demand leads to a larger increase in equilibrium output because a further increase in consumption demand is then induced. We explain below why this is called the *multiplier*.

Desired saving is $S = Y - C = Y - [A + cY] = -A + (1 - c)Y$, the saving function corresponding to the consumption function $C = A + cY$. In equilibrium, equation (1) implies

$$I = Y^*(1 - c) - A \quad (2)$$

But the right-hand side of equation (2) is simply desired saving in equilibrium when output is Y^* . Hence in equilibrium $I = S$. Planned investment equals planned leakages.

Planned versus actual

Equilibrium output and income satisfy two equivalent conditions. Aggregate demand must equal income and output. Equivalently, planned investment must equal planned saving.

In the previous chapter we showed that *actual* investment is *always* equal to *actual* saving, purely as a consequence of our national income accounting definitions. When the economy is not in equilibrium, planned saving and investment are not equal. However, unplanned investment in stocks and/or unplanned saving (frustrated consumers) always ensures that actual investment, planned plus unplanned, equals actual saving, planned plus unplanned.

16.5 A fall in aggregate demand

The *slope* of the *AD* schedule depends only on the marginal propensity to consume (MPC). For a given MPC, the level of autonomous spending [$A + I$] determines the *height* of the *AD* schedule. Autonomous spending is spending unrelated to income.

Changes in autonomous spending lead to parallel shifts in the *AD* schedule. Investment demand depends chiefly on current guesses by firms about future demand for their output. Beliefs about this future demand can fluctuate significantly, influenced by current pessimism or optimism about the future. Similarly, a fall in consumer confidence reduces autonomous consumption demand.

Suppose firms get pessimistic about future demand for their output. Planned investment falls. If autonomous consumption is unaffected, the

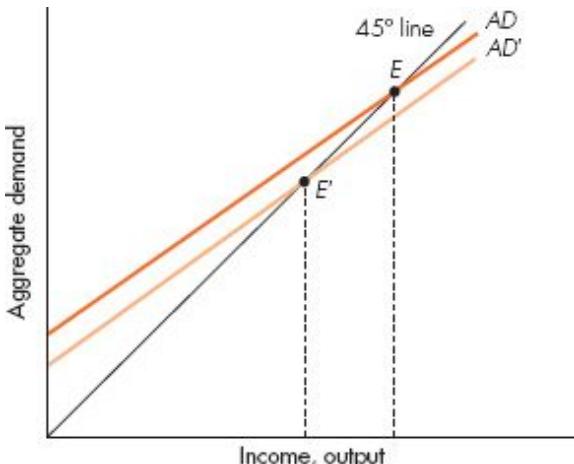
aggregate demand schedule AD is now lower at each income than before. Figure 16.8 shows this downward shift from AD to AD' .

Before we go into the details, think about what is likely to happen to output. It will fall, but how much? When investment demand falls, firms cut output. Households have lower incomes and cut consumption. Firms cut output again, further reducing household incomes. Consumption demand falls further. What brings the process of falling output and income to an end?

Figure 16.8 shows that a given downward shift of the AD schedule reduces equilibrium output by a *finite* amount, but by an amount larger than the vertical fall in the AD schedule. This is because the AD schedule has a slope flatter than the 45-degree line: its slope, the marginal propensity to consume, is always smaller than unity.

Equilibrium moves from E to E' . Equilibrium output falls *more* than the original cut in investment demand, but does not fall all the way to zero.

Table 16.1 explains. Since many students find arithmetic easier than algebra, we illustrate for the particular values [$A = 10$] for autonomous consumption demand and [$c = 0.6$] for the marginal propensity to consume. Thus the consumption function is $C = 10 + 0.6Y$.



When investment demand falls, the aggregate demand schedule shifts down from AD to AD' and equilibrium output falls by a larger *amount*.

Figure 16.8 A fall in investment demand

Table 16.1 Adjustment to a shift in investment demand

	Y	I	$C = 10 + 0.6Y$	$AD = C + I$	$Y - AD$	Unplanned stocks	Output
Step 1	100	30	70	100	0	Zero	Constant
Step 2	100	20	70	90	10	Rising	Falling
Step 3	90	20	64	84	6	Rising	Falling
Step 4	84	20	60.4	80.4	3.6	Rising	Falling
New equilibrium	75	20	55	75	0	Zero	Constant

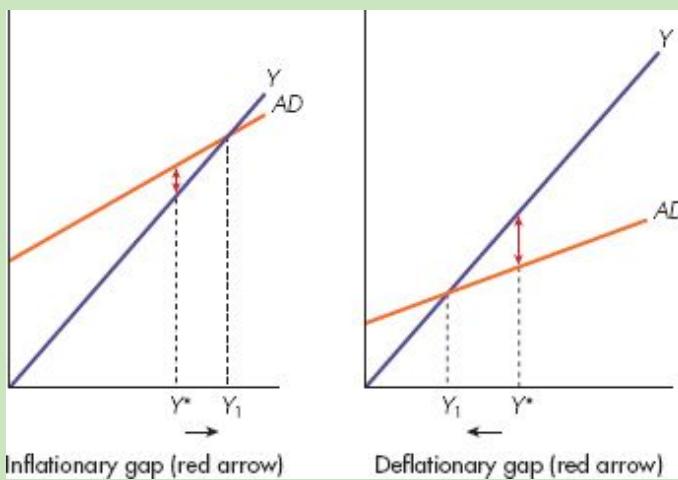
If original investment demand is 30, the first row of Table 16.1 shows that the original equilibrium output is 100, since consumption demand is then $[10 + 60]$ and investment demand is 30. Thus aggregate demand just equals actual output.

In step 2, investment demand falls to 20. Firms did not expect demand to change, and still produced 100. Output exceeds aggregate demand by 10. Firms add this 10 to unplanned inventories, then cut output to get rid of these again.

CONCEPT 16.2

INFLATIONARY AND DEFLATIONARY GAPS

We now make explicit the extent to which short-run equilibrium output deviates from potential output. First, we need some terminology. Nowadays, we use inflation and deflation to refer to rises and falls in the price level. Earlier economists assumed that these were largely caused by output levels that were above or below potential output. We now know that this is one cause, but not the only cause, of price changes, as we discuss in later chapters.



Even so, some of the old terminology survives. The **inflationary gap** is a measure of the extent to which aggregate demand exceeds the level needed to ensure output equals potential output, and the **deflationary gap** is a measure of the extent to which aggregate demand falls short of the level required to achieve potential output.

Both figures above show aggregate demand AD and the 45-degree line along which demand equals output Y . In the left-hand diagram, aggregate demand exceeds that at which equilibrium output Y and potential output Y^* would coincide. We can measure this excess either in the vertical direction or in the horizontal direction.

Thus, in the left-hand figure, the inflationary gap corresponds to the vertical distance shown in red. Conversely, in the right-hand figure, aggregate demand is insufficient to allow equilibrium output Y_1 and potential output Y^* to coincide.

Thus, in the right-hand figure, the deflationary gap corresponds to the vertical distance shown in red.

We could instead measure these excesses and shortfalls in the horizontal direction rather than the vertical direction.

The **inflationary gap** measures the excess of aggregate demand over output when output is at potential output.

The **deflationary gap** measures the shortfall of aggregate demand over output when output is at potential output.

The **output gap** measures the discrepancy between output and potential output.

Thus, in the left-hand figure, the **output gap** is the positive amount shown by the black arrow, the distance between Y_1 and Y^* . In the right-hand figure, the output gap is the negative distance shown by the black arrow, the amount by which Y_1 falls short of Y^* .

Modern economics tends to use output gaps rather than inflationary and deflationary gaps to describe the strength of aggregate demand. We will show empirical estimates of output gaps in later chapters.

Step 3 shows firms making 90, the level of demand in step 2. But when firms cut output, income falls. Step 3 shows consumption demand falls from 70 to 64. Since the *MPC* is 0.6, a cut in income by 10 has caused a fall in consumption demand by 6. The induced fall in consumption demand means that output of 90 still exceeds aggregate demand, which is now 84. Again inventories pile up unexpectedly, and again firms respond by cutting output.

At step 4, firms make enough to meet demand at step 3. Output is 84, but again this induces a further cut in consumption demand. Output still exceeds aggregate demand. The process keeps going, through many steps, until it reaches the new equilibrium, an output of 75. Output and income have fallen by 25, consumption demand has fallen by 15 and investment demand has fallen by 10. Aggregate demand again equals output.

How long it takes for the economy to reach the new equilibrium depends on how well firms figure out what is going on. If they keep setting output targets to meet the level of demand in the previous period, it takes a long time to adjust. Smart firms may spot that, period after period, they are overproducing and adding to unwanted inventories. They anticipate that demand is still falling and cut back output more quickly than Table 16.1 suggests.

Why does a fall of 10 in investment demand cause a fall of 25 in equilibrium output? Lower investment demand induces a cut in output and income that then induces an extra cut in consumption demand. Total demand falls by more than the original fall in investment demand, but the process does not spiral out of control. Equilibrium output is 75.

The **multiplier** is the ratio of the change in equilibrium output to the change in autonomous spending that caused the change.

In our example, the initial change in autonomous investment demand is 10 and the final change in equilibrium output is 25. The **multiplier** is 2.5. That is why, in Figure 16.7, a small downward shift in the *AD* schedule leads to a much larger fall in equilibrium income and output.

16.6 The multiplier

The multiplier tells us how much output changes after a shift in aggregate demand. The multiplier exceeds 1 because a change in autonomous

demand sets off further changes in consumption demand. The size of the multiplier depends on the marginal propensity to consume. The initial effect of a unit fall in investment demand is to cut output and income by a unit. If the MPC is large, this fall in income leads to a large fall in consumption and the multiplier is big. If the MPC is small, a given change in investment demand and output induces small changes in consumption demand and the multiplier is small.

Table 16.2 examines a one-unit increase in investment demand. In step 2, firms raise output by 1 unit. Consumption rises by 0.6, the marginal propensity to consume times the one-unit change in income and output. At step 3, firms raise output by 0.6 to meet the increased consumption demand in step 2. In turn, consumption demand is increased by 0.36 (the MPC 0.9 multiplied by the 0.9 increase in income) leading in step 4 to a rise in output of 0.36. Consumption rises again and the process continues.

Table 16.2 Calculating the multiplier when the MPC equals 0.6

Change in	Step 1	Step 2	Step 3	Step 4	Step 5	*	*	*
I	1	0	0	0	0	*	*	*
Y	0	1	0.6	$ 0.6 ^2$	$ 0.6 ^3$	*	*	*
C	0	0.6	$ 0.6 ^2$	$ 0.6 ^2$	$ 0.6 ^3$	*	*	*

To find the multiplier, we add all the increases in output from each step in the table and keep going:

$$\text{Multiplier} = 1 + (0.6) + (0.6)^2 + (0.6)^3 + (0.6)^4 + (0.6)^5 + \dots$$

The dots at the end mean that we keep adding terms such as $(0.9)^6$ and so on. The right-hand side of this equation is called a geometric series. Each term is (0.9) times the previous term. Fortunately, mathematicians have shown that there is a general formula for the sum of all the terms in such a series:

$$\text{Multiplier} = 1/(1 - 0.6) = 1/0.4 = 2.5$$

The formula applies whatever the (constant) value of c , the marginal propensity to consume:

$$\text{Multiplier} = 1 + c + c^2 + c^3 + c^4 + c^5 + \dots = 1/(1 - c)$$

For the particular value of $c = 0.6$, the multiplier is $1/(0.4) = 2.5$. Hence a cut in investment demand by 10 causes a fall in equilibrium output by 25, as we know from Table 16.1. For those of you who 'did the maths' above,

equilibrium output is simply autonomous demand multiplied by the multiplier!

As an example, suppose $c = \frac{1}{2}$. The multiplier is then $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$. You can probably guess that this is eventually going to cumulate to 2, which is exactly what the general formula promises us.

The marginal propensity to consume tells how much of each extra unit of income is spent on consumption. Thus c is a number between zero and unity. The higher is c , the lower is $(1 - c)$. Dividing 1 by a smaller number leads to a larger answer. The general formula for the multiplier confirms that a larger c implies a larger multiplier.

The multiplier and the *MPS*

Any part of an extra unit of income not spent must be saved. Hence $1 = c + s$, where s is the marginal propensity to save.

Hence we can also think of the multiplier as $1/s$. The higher the **marginal propensity to save**, the more of each extra unit of income leaks out of the circular flow into savings and the less goes back round the circular flow to generate further increases in aggregate demand, output and income. Since the marginal propensity to save is a positive fraction, the multiplier exceeds unity, as we already know.

The **marginal propensity to save** is the fraction of each extra unit of income that households wish to save.

In the next chapter, we will see that, after introducing the government and foreign sectors, it remains true that the multiplier equals $1/mpl$ where mpl is the marginal propensity to leak out of the circular flow. Here, saving is the only leakage. Whatever the leakages, a 1-unit increase in injections (autonomous demand), from whatever source, must eventually induce a 1-unit increase in desired leakages to restore equilibrium. Once we know the marginal propensity to leak, we know how much output has to rise to create a 1-unit increase in leakages. In this chapter, the only leakage is through saving, so the multiplier is $1/s$.

$1 =$ Assumed rise in autonomous demand

$=$ Rise in planned injections

- =Rise in planned leakages once equilibrium restored
- = (Rise in output, ΔY) x (marginal propensity to save, s) Hence, $1 = \Delta Y \times s$, so $\Delta Y = 1/s$.

16.7 The paradox of thrift

The previous section analysed a change in equilibrium output caused by a change in autonomous investment demand. We now examine the consequences of a change in the autonomous part of planned consumption and saving.

We could use the 45-degree diagram to show how aggregate demand shifts and causes a change in equilibrium output. Suppose households increase autonomous consumption demand by 10. There is a parallel upward shift in the consumption function, and hence also in the aggregate demand schedule AD . A higher aggregate demand schedule must intersect the 45-degree line at a higher level of output. Hence, equilibrium output increases.

But what happens to planned saving? Households wish to consume more (save less) at each level of income, but now face a higher equilibrium income. They save a lower fraction of a higher income. Does saving rise or fall? In equilibrium, planned saving *always* equals planned investment, and the latter is unaltered. Hence planned saving cannot change. Equilibrium income must therefore have risen just enough to offset the desire to save a lower fraction of any particular income level, leaving overall planned saving unaltered once the new equilibrium is reached.

This chain of reasoning may seem quite complicated. Fortunately, it can be grasped much more directly, using Figure 16.9, which focuses on the equality of planned saving and planned investment in equilibrium. A decline in thriftiness - a fall in the desire to save – shifts the planned saving schedule from S to S' . Planned investment is unaffected. Hence equilibrium income must rise from Y^* to Y^{**} to maintain the equality of planned saving and planned investment.

A change in the amount households wish to save at each income leads to a change in equilibrium income, but *no change* in equilibrium saving, which must still equal planned investment. This is the **paradox of thrift**.

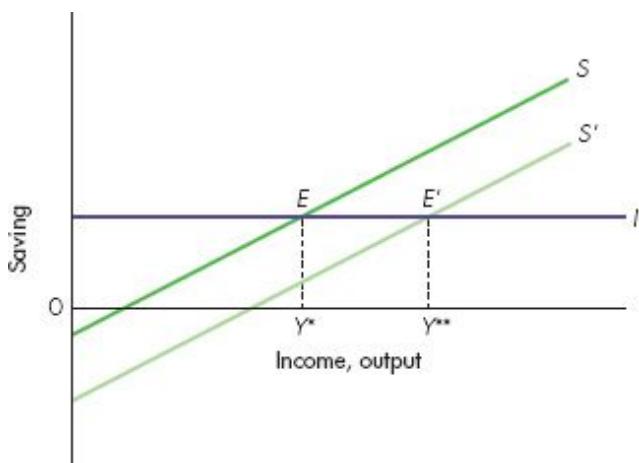
The **paradox of thrift** helps us to understand an old debate about the virtues of saving and spending. Does society benefit from thriftiness and a high level of desired saving at each income level? The answer depends on whether or not the economy is at full employment.

When aggregate demand is low and the economy has spare resources, the paradox of thrift shows that a *reduction* in the desire to save (higher desire to consume) will increase aggregate demand and the equilibrium output level. Society benefits from higher output and employment. Since investment demand is autonomous, a change in the desire to save has no effect on the desired level of investment.

Applying the same argument in reverse, during the crash of 2008 politicians were worried that the panic might lead to too great a desire to save. At a time when aggregate demand had already fallen, equilibrium output had fallen even with a constant propensity to save: any additional desired saving would reduce equilibrium output even further. Case 16.4 discusses the saving rate in more detail.

In contrast, think what happens in the long run once the economy has time to return to the level of potential output (see Chapter 21). If output equals potential output, an *increase* in the desire to save will increase saving, and reduce consumption, at potential output. Other adjustments may then induce investment demand to rise to restore aggregate demand to its full-employment level. The next few chapters explain why. Hence, in the long run, society may benefit from an *increase* in the desire to save. Investment will rise and the economy's capital stock and potential output may grow more quickly.

In this chapter we have focused on the short run before prices and wages have time to adjust. Saving and investment decisions are made by different people. There is no automatic mechanism to translate higher saving into a corresponding rise in investment demand. Since planned saving depends on the level of income, income adjusts to equate planned saving and planned investment.



In equilibrium planned saving equals planned investment. A fall in the desire to save induces a rise in equilibrium output to keep planned saving equal to planned investment.

Figure 16.9 The paradox of thrift

CASE 16.4

HOW STABLE IS THE SAVING RATE?

The saving rate S/Y is the fraction of income that is saved. In the Thatcher boom of the late 1980s, heady optimism and easy access to credit made UK consumers spend a lot. Personal saving collapsed as people bought champagne, sports cars and houses. The boom didn't last. As inflation rose, the government raised interest rates to slow down the economy. House prices fell. People's mortgage debt was larger than the value of their houses. To pay off this 'negative equity', households raised saving sharply in the early 1990s.

During 1992–2008, UK households were borrowing again. Low interest rates fuelled a spending boom and a protracted rise in house prices. People saved less and borrowed more in order to spend. TV shows such as *The Property Ladder* and *Location, Location, Location* showed people how to renovate houses for subsequent letting or sale. In a rising market, people made money on buying and selling houses whether or not they were actually any good at redeveloping them.

The chart shows household saving, as a percentage of household disposable income. It shows the saving rate fell sharply in the spending boom of the late 1980s, rose in the recession of the early 1990s, then fell steadily during the sustained boom of 1992–2008. By 2008, UK households were saving less than 2 per cent of their disposable income. With steady economic growth assured, people thought it made sense to borrow rather than save. The good times had arrived.



UK household saving (% of disposable income), 1983-2012

Source: OECD. Economic Outlook No. 93, 2013.

What do you think happened as a result of the financial crash in 2008? People got scared. Property prices began to fall and borrowing for house purchase no longer seemed a good idea. Banks became terrified their customers could not repay, so the supply of new lending from banks dried up completely. Credit card debt suddenly seemed expensive. And people who foresaw a deep recession began to cut out unnecessary expenditure. They chose to save a larger fraction of their income.

The saving rate rose very sharply in 2009, as the chart confirms. The saving rate is likely to stay high until (a) households feel they have paid off a reasonable amount of their large inherited debts and (b) confidence in future growth of income and employment returns.

Clearly, then, the saving rate can fluctuate a lot. Although in this chapter we assume a constant marginal propensity to save, Chapter 20 discusses more sophisticated theories of consumption and saving that help us understand why the saving rate fluctuates.

One final remark. Does it matter whether households borrow in order to buy a foreign holiday or to buy a house for subsequent rental to others? In the former case, no asset is purchased for the future; in the latter case, the household acquires an asset that will give rise to future incomes. Simply measuring today's income and today's spending gives a misleading picture of the long-run economic position of the household. We return to this issue in Chapter 19.

16.8 The role of confidence

Shifts in autonomous demand - whether autonomous consumption demand or autonomous investment demand - are often caused by changes in confidence; that is, swings in optimism or pessimism about the future. Our simple model assumes these are independent of current income and output.

This does not mean that they are not important, and not subject to influence by policy and politicians. Rather than spend taxpayers' money trying to boost aggregate demand through a subsidized car scrappage scheme, most governments would rather talk up demand if only they could.

Like the boy who cried wolf once too often, governments who mislead the public soon become distrusted, and their warm words are then ignored. However, if governments can provide a clear and credible account of why the future may be rosier than the present, they may indeed be able to stimulate aggregate demand through increasing confidence and thereby inducing households and firms to spend more. Conversely, when they announce bad news that had not previously been foreseen, then at any particular level of current output firms and households will reduce their demand, and aggregate demand will fall.

CONCEPT 16.3

SAVING OR SAVINGS?

What is the difference between saving and savings? Can we use them interchangeably? Should we talk about the savings rate or the saving rate?

To get things right, we need to return to the distinction between stocks and flows. A stock can be measured at a point in time, whereas a flow needs a time dimension and the size of the flow vanishes as the time interval becomes shorter and shorter. The volume of water in a basin is a stock, still existing if we take a picture with a high-speed camera. The inflow from the tap or outflow down the plughole are flows - if we contemplate a short enough nano-second, essentially there is no time for water to flow in or out.

So far, our national accounts have been about flows: the flow of output per year, the flow of consumption spending or investment spending during that period. If we think about a month, rather than a year, we get a different numerical answer for the size of GDP or consumption. Corresponding to these flows, in principle there are a set of stock accounts. For the country, this is the stock of national assets, the stock of national liabilities, and hence net national wealth. Similarly, we can think about the assets and liabilities of households or of firms.

Saving is the flow of income not spent within a period. It is always a flow concept. *Savings* is the stock of assets accumulated as a result of the past flows of saving decisions.

That is why it is correct to talk about the saving rate, but wrong to talk about the savings rate. If we want to examine how much of the flow of disposable income is not spent, we need to use the flow measure, *saving*, not the stock measure, *savings*.

Summary

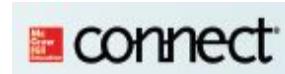
- **Aggregate demand** is planned spending on goods (and services). The *AD* schedule shows aggregate demand at each level of income and output.
- This chapter neglects planned spending by foreigners and by the government and studies **consumption demand** by households and **investment demand** by firms (desired additions to physical capital and to inventories). We treat investment demand as constant.

- Consumption demand is closely, though not perfectly, related to **personal disposable income**. Without taxes or transfers, personal disposable income and total income coincide.
- **Autonomous consumption** is desired consumption at zero income. The **marginal propensity to consume (MPC)** is the fraction by which planned consumption rises when income rises by a pound. The **marginal propensity to save (MPS)** is the fraction of an extra pound of income that is saved. Since income is consumed or saved, $MPC + MPS = 1$.
- For given prices and wages, the goods market is in equilibrium when output equals planned spending or aggregate demand. Equivalently, in equilibrium, planned saving equals planned investment. Goods **market equilibrium** does not mean output equals potential output. It means planned spending equals actual spending and actual output.
- The **equilibrium output is demand-determined** because we assume that prices and wages are fixed at a level that implies an excess supply of goods and labour. Firms and workers are happy to supply whatever output and employment is demanded.
- When aggregate demand exceeds actual output there is either unplanned disinvestment (inventory reductions) or unplanned saving (frustrated customers). Actual investment always equals actual saving, as a matter of definition. Unplanned inventory reductions or frustrated customers act as a signal to firms to raise output when aggregate demand exceeds actual output. Similarly, unplanned additions to stocks occur when aggregate demand is below output.
- A rise in planned investment increases equilibrium output by a larger amount. The initial increase in income to meet investment demand leads to further increases in consumption demand.
- The **multiplier** is the ratio of the change in output to the change in autonomous demand that caused it. In the simple model of this chapter,

the multiplier is $1/[(1 - MPC)]$ or $1/MPS$. The multiplier exceeds 1 because MPC and MPS are positive fractions.

- The **paradox of thrift** shows that a reduced desire to save leads to an increase in output but no change in the equilibrium level of planned saving, which must still equal planned investment.

Review questions



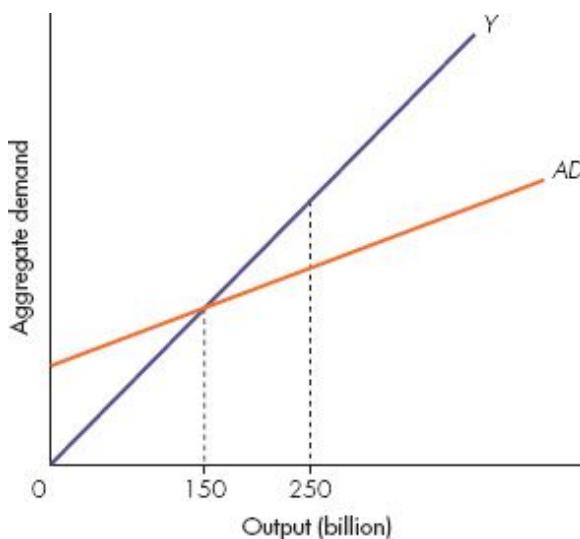
EASY

- 1 (a) Find equilibrium income when investment demand is 400 and $C = 0.8Y$. (b) Would output be higher or lower if the consumption function were $C = 100 + 0.7Y$?
- 2 Which of the following statements is correct? (a) Any tax is a tax on jobs because it reduces aggregate demand. (b) Provided the government spends the tax revenue, the impact of higher spending outweighs the adverse demand effect of higher taxes. (c) Autonomous consumption demand is directly related to consumer confidence. (d) All the above statements could be true, depending on the other things assumed equal.
- 3 Suppose firms are initially surprised by changes in demand. (a) When demand falls, what is the initial effect on stocks of unsold goods held by firms? (b) What do firms plan to do to stocks as soon as they have time to adjust production? Does this reduce or increase the initial fall in demand? (c) Once stocks have been adjusted, what then happens to production and output?
- 4 Could the multiplier ever be less than 1?
- 5 **Common fallacies** Why are these statements wrong? (a) If people were prepared to save more, investment would increase and we could get the economy moving again. (b) Lower output leads to lower spending and yet lower output. The economy could spiral downwards forever.

MEDIUM

- 6 Assume that an economy is in equilibrium. Planned investment is £100. The MPC is 0.6. Suppose investment rises by £30. (a) What happens to the equilibrium output?
Now suppose people decide to save a higher proportion of their income: the consumption function changes from $C = 0.8Y$ to $C = 0.5Y$. (b) What happens to equilibrium income (planned investment being £100)?

- (c) What happens to the equilibrium proportion of income saved? Explain.
- 7 Assume that the economy is in equilibrium. The MPC is 0.6. Suppose investment demand rises by £30. (a) By how much does the equilibrium output increase? (b) How much of that increase is extra consumption demand? Draw the corresponding diagram using planned investment and planned saving assuming that the initial output is 100.
- 8 Suppose the consumption function is $C = 0.75 Y$ and planned investment is 40. (a) Draw a diagram showing the aggregate demand schedule. (b) If actual output is 100, what unplanned actions will occur? (c) What is the equilibrium output? (d) Do you get the same answer using planned saving equals planned investment?
- 9 The diagram below shows the aggregate demand and output of an economy. Along the 45-degree line Y , demand equals output. Is there an inflationary gap or a deflationary gap if the answer is 250?
- 10 Suppose confidence depends a little on the current level of output, and the model therefore becomes $I = aY + I^* C = [A] + cY = [A^* + BY] + cY$ where I^* and A^* remain autonomous and independent of output, but a and b reflect the dependence of confidence on the current level of output. (a) What is the new value of the multiplier? (b) Is this higher or lower than before? (c) Is equilibrium output higher or lower than before?



- 11 When could the paradox of thrift fail to be true?
- HARD
- 12 **Essay question** 'The remarkably strong relationship between consumption and income confirms that most people want to spend most

of their income as soon as they can. We are all material girls and boys at heart.' Is the inference justified?

- |3 Planned investment is 100. Initially, the consumption function is $C = 100 + 0.8 Y$. There are three ways in which greater pessimism about the future might affect behaviour: (a) planned investment falls from 100 to 50, (b) autonomous consumption falls from 100 to 50, (c) the marginal propensity to consume falls from 0.8 to 0.7 as people save more of each unit of additional income. Draw a graph of each change and its effect on short-run equilibrium output.
 - |4 Suppose your economy is going through a recession. Individuals desire to save more and spend less. How does the paradox of thrift explain the consequences of increased savings in your economy?
- 1 A is the minimum consumption needed for survival. How do households finance it when their incomes are zero? In the short run they dissave and run down their assets. But they cannot do so for ever. The consumption function may differ in the short run and the long run, an idea we discuss in Chapter 20.

CHAPTER 17

Fiscal policy and foreign trade

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 analyse how fiscal policy affects aggregate demand
- 2 discuss equilibrium output in this extended model
- 3 calculate the balanced budget multiplier
- 4 evaluate automatic stabilizers
- 5 explain the structural budget and the inflation-adjusted budget
- 6 discuss how budget deficits add to national debt
- 7 evaluate the limits to discretionary fiscal policy
- 8 define imports and exports
- 9 analyse how foreign trade affects equilibrium output

The previous chapter discussed aggregate demand in a model with only firms and households. Now we need to reintroduce the other two sectors, the government and countries overseas. This chapter examines aggregate demand and output determination in this expanded model.

17.1 The scope of government activity

In most European countries, the government directly buys about a fifth of national output and spends about the same again on transfer payments. This spending is financed mainly by taxes, though some is financed by borrowing. What is the macroeconomic impact of government **fiscal policy**? Why did governments conclude that a massive fiscal response was required when confronted with the financial crash of 2008? Has the

extra government debt then incurred constrained the subsequent use of fiscal policy to boost demand and economic recovery?

We first show how fiscal policy affects aggregate demand and equilibrium output. Then we study three fiscal issues. We analyse opportunities and limitations in using fiscal policy to **stabilize output**.

We then examine the significance of the government's **budget deficit**. When the government runs a deficit, it spends more than it earns. How can the government keep spending more than it receives? We examine the size of the deficit, and ask how much it matters.

A government deficit is financed mainly by borrowing from the public by selling bonds, which are promises to pay specified amounts of interest payments at future dates in exchange for cash up front. This borrowing adds to **national debt** to the public.¹ During 2009 governments around the world had huge budget deficits as they bailed out their banking systems and spent money on car scrappage schemes to try to prevent the car industries imploding. Just as for an individual, when a government spends more than it earns it adds to its debts.

Fiscal policy is government policy on spending and taxes. It affects the size of government deficits and thus government debt.

Stabilization policy is government action to keep aggregate demand and actual output close to potential output.

The **budget deficit** is the excess of government spending over government receipts during a particular period.

The **national debt** is the stock of government debt outstanding.

Most of this chapter is about the government's role in aggregate demand, but to complete our model of output determination, we must also add foreign trade. Exports X and imports Z are each about 15 per cent of GDP in a large country such as the United States (which mainly trades with itself), but can reach up to 75 per cent of GDP in a small open economy such as Belgium or the Netherlands. In middle-sized countries such as the UK, France and Germany, exports and imports are around 30 per cent of GDP. Thus, the effects of foreign trade are too important to ignore.² Exports add to aggregate demand for domestic output, but

imports reduce aggregate demand for domestic output by diverting desired spending to foreign output.

17.2

Government and aggregate demand

Government spending G on goods and services adds directly to aggregate demand. The government also withdraws money from the circular flow through indirect taxes T^e on expenditure and direct taxes T^d on factor incomes, less transfer benefits B that augment factor incomes. However, transfer payments affect aggregate demand only by affecting other components such as consumption or investment demand.

Since it is a pain to keep distinguishing between market prices and basic prices, we assume all taxes are direct taxes. With no indirect taxes, measurements at market prices and at basic prices coincide. Initially, we still ignore foreign trade.

Aggregate demand AD is consumption demand C , investment demand I and government demand G for goods and services. Transfer payments affect aggregate demand only by affecting C or I . It would be double-counting to include transfer payments directly in aggregate demand. Thus $AD = C + I + G$.

In the short run, we assume G is fixed, or at least independent of income. Its size reflects how many hospitals the government wants to build and how many teachers it wants to hire. We now have three autonomous components of aggregate demand independent of current income and output: autonomous consumption demand, investment demand I and government demand G .

The government also levies taxes and pays out transfer benefits. **Net taxes** NT are taxes minus transfer benefits. Net taxes reduce personal disposable income – the amount available for spending or saving by households – relative to national income and output. If YD is **disposable income**, Y national income and t the *net taxrate* (assumed to be a constant proportion of income), then disposable income $YD = (1 - t)Y$.

Net taxes are taxes minus transfer benefits.

Disposable income is gross income minus taxes plus benefits; that is, the net income available to spend or save.

If taxes, net of transfer benefits, are 20 per cent of national income, the net tax rate t is 0.2. If national income rises by £1, net tax revenue rises by 20 pence, and household disposable income rises by 80 pence.

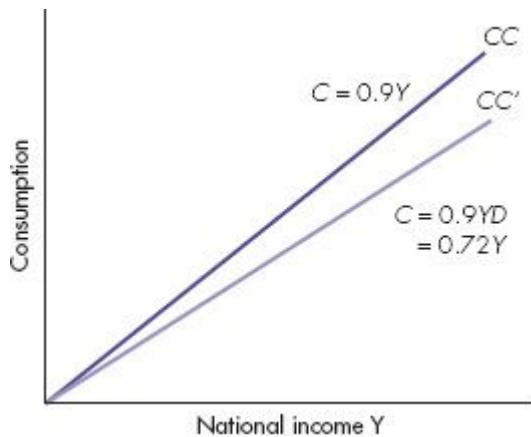
Suppose households wish to consume 90 per cent of their *disposable* income. Since disposable income is only 80 per cent of national income, the marginal propensity to consume out of national income is only 0.72. More generally, if the marginal propensity to consume out of disposable income is c , the marginal propensity to consume out of national income c' is reduced to $c(1-t)$.

The larger the net tax rate, the smaller is c' and the flatter is the slope of the consumption function, when plotted against national income and output. Figure 17.1 illustrates.

The effect of net taxes on output

Suppose initially that government spending is zero. Figure 17.2 illustrates. A rise in the net tax rate from zero to 0.2 made the consumption function pivot downwards from CC to CC' in Figure 17.1. Adding on the constant investment demand I to that consumption function, we obtain aggregate demand. Hence, a rise in the net tax rate rotates the consumption function from CC to CC' in Figure 17.1 and rotates aggregate demand from AD to AD' in Figure 17.2. Aggregate demand equals actual output at a lower output level than before, at E' not E . Equilibrium income and output are lower.

Raising the net tax rate reduces equilibrium output. Conversely, if aggregate demand and equilibrium output are below potential output, lower tax rates or higher transfer benefits will raise aggregate demand and hence equilibrium output.



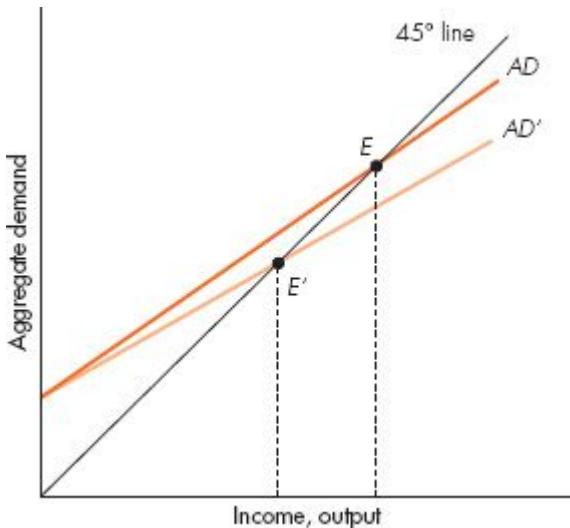
In the absence of taxation, national income Y and disposable income YD are the same. The consumption function CC' shows how much households wish to consume at each level of national income. With a proportional net tax rate of 0.2, households still consume 90p of each pound of disposable income. Since YD is now only $0.8Y$, households consume only $0.9 \times 0.8 = 0.72$ of each extra unit of national income. Relating consumption to national income, the effect of net taxes is to rotate the consumption function downwards from CC to CC' .

Figure 17.1 Net taxes and consumption

The effect of government spending on output

Now forget taxes and think government spending. Suppose the net tax rate is zero. National income and disposable income once again coincide. Figure 17.3 shows that a higher level of autonomous government spending has an effect similar to that of higher autonomous investment demand studied in Chapter 16. With a marginal propensity to consume of 0.9, the multiplier is again $1/(1 - c) = 10$.

A rise in government spending G induces a rise in equilibrium output by 10 times that amount. In Figure 17.3 equilibrium moves from E to E' as aggregate demand shifts from AD to AD' .



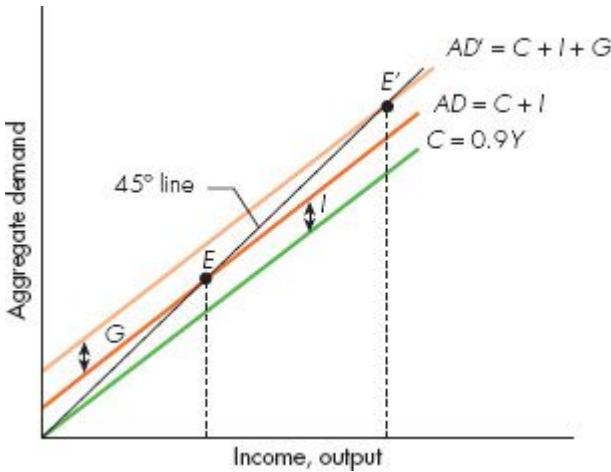
An increase in the income tax rate or a reduction in rate of unemployment benefit will increase the net tax rate t . The consumption function rotates from CC to CC' in Figure 17.1. With constant investment demand, the aggregate demand schedule rotates from AD to AD' in Figure 17.2. The equilibrium level of output falls and the equilibrium point moves from E to E' .

Figure 17.2 A higher net tax rate

The combined effects of government spending and taxation

Suppose an economy begins with equilibrium output of 1000 but no government. Assume demand from autonomous consumption and investment is 100. With a marginal propensity to consume out of disposable income of 0.9, a disposable income of 1000 induces consumption demand of 900. Aggregate demand is $(900 + 100) = 1000$, which is also actual output.

Now add autonomous demand of 200 from the government, taking total autonomous demand to 300. Also introduce a net tax rate of 0.2. The marginal propensity to consume out of national income falls from 0.9 to 0.72, and the multiplier becomes $1/(1 - 0.72) = 1/0.28 = 3.57$. Multiplying autonomous demand of 300 by 3.57 yields equilibrium output of 1071, above the original equilibrium output of 1000. Figure 17.4 illustrates.



Beginning from equilibrium at E, AD shifts up to AD'. The new equilibrium is E' and equilibrium output rises by 10 times the rise in G.

Figure 17.3 Government spending and equilibrium output

The balanced budget multiplier

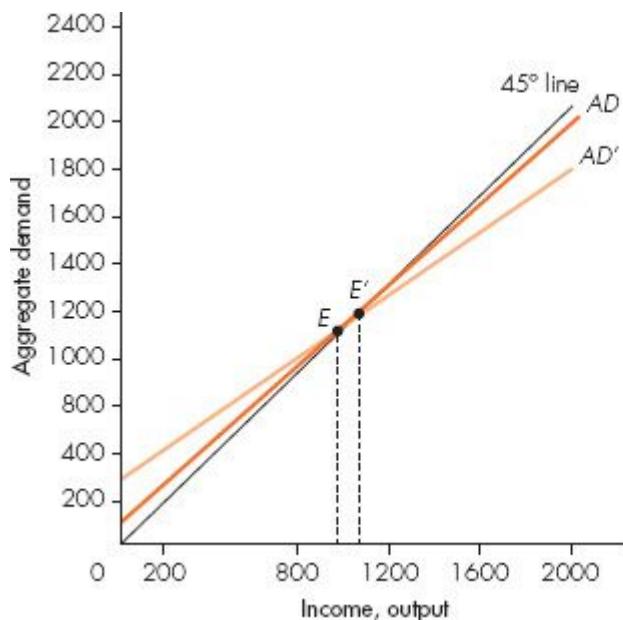
The economy began at an equilibrium output of 1000. With a proportional tax rate of 20 per cent, initial tax revenue was 200, precisely the amount of government spending.

This balanced increase in government spending and taxes did not leave demand and output unaltered. Figure 17.4 shows equilibrium output is larger. The new 200 of government spending raises aggregate demand by 200. The tax increase cuts disposable income by 200, but with $c = 0.9$ lower disposable income reduces consumption demand by only 180.

The initial net effect of the tax and spending package raises aggregate demand by 20. Output rises, inducing further rises in consumption demand. When the new equilibrium is reached, output has increased by 71, from 1000 to 1071. This is the famous **balanced budget multiplier**, which gives the government a fiscal tool to boost aggregate demand without adding to the deficit or debt.

The **balanced budget multiplier** says that a rise in government spending plus an equal rise in taxes leads to higher output.

To use this tool, however, the government has to have the political courage to raise tax revenue in line with higher expenditure. Sometimes governments are unable or unwilling to do this.



Beginning from equilibrium at E, government spending rises from zero to 200, shifting the AD schedule upwards, and the tax rises from zero to 0.2, making the new schedule AD'. Equilibrium moves from E to E' where AD' intersects the 45 line. Equilibrium output increases from 1000 to 1071.

Figure 17.4 Higher spending and taxes

CASE 17.1

FISCAL POLICY, AUSTERITY AND DEBT: LESSONS FROM JAPAN

For the last 20 years, this textbook has used the Japanese example both to illustrate what hypothetically might happen to Western economies if they got into a macroeconomic mess, and to draw lessons for what policy options were available. Since 2009, Western economies have been in that mess themselves. The Japanese example is more relevant than ever.

After three decades of post-war success, Japanese economic growth came to an abrupt end in the 1990s. A crash in property prices made banks bankrupt. Instead of admitting this and sorting it out, policy makers ignored the problem. Consumers lost confidence, and output fell. To restore confidence, Japan had big fiscal expansions

to boost demand. Why did Japanese output not recover in the 1990s?

Facing fiscal expansion in a severe recession, Japanese households and firms decided aggressively expansionary government policy was being undertaken only because the government knew things were even worse than the private sector had previously thought. The private sector took on board this new information and became even more pessimistic. The autonomous parts of consumption and investment demand fell sufficiently to offset the fiscal expansion injected by the government; and this fall in autonomous demand was therefore caused by the expansionary policy itself. Fiscal expansion failed to boost output by much. In macroeconomics the induced effects can outweigh the direct effect. Not until 2010 did sustainable growth appear to return.

After further bouts of fiscal stimulus, by 2013 Japan's government net debt had reached nearly 140 per cent of GDP, the highest among developed economies. Worries about how this debt will be financed in the future also undermine the confidence of households and firms, further reducing consumption and investment demand. The table below illustrates Japan's economic misery during this period.

Japan's macroeconomic misery 1993–2012

	Annual GDP growth (%)	Interest rate (%)	Budget deficit (% of GDP)	Government net debt (% of GDP)
1991–92	2.5	6.5	-1.5	13
1993–94	0.5	2.5	3.0	19
1994–95	1.5	1.5	3.0	23
1995–96	2.5	2.5	5.0	28
1997–99	-1.0	0.5	7.5	34
2000–02	1.0	0	6.0	67
2003–07	1.5	0	4.5	81
2008–09	-3.5	0	5.0	100
2010–12	2.0	0	9.5	134

Source: Based on data from Economic Outlook Statistical Annex, various issues © OECD, accessed on 18/06/2013.

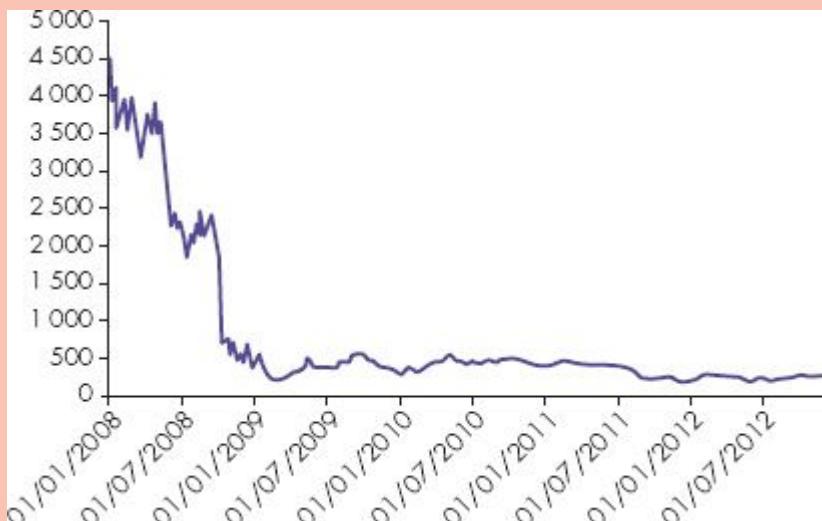
Cutting interest rates to near zero levels probably prevented the outcome being even worse, but alone it was insufficient to restore healthy growth. Japan's government deficits ranged from 3 to 9.5 per cent during the 20-year period from 1993–2012 as it attempted

fiscal stimulus to boost aggregate demand. Again, this may have prevented something even worse, but it was certainly not sufficient to resolve the crisis; other components of autonomous demand fell. And it has now left Japan with a massive level of government debt.

Lessons for Western economies

The Japanese example thus contains three important lessons. First, when confidence collapses, even fiscal policy may not be able to boost aggregate demand. This means that governments should do all they can to prevent confidence ever collapsing to this extent. In retrospect, Western banks were too loosely regulated during the long boom before 2007. Governments had failed to discharge their responsibility to create a stable financial environment.

Second, cleaning up the banks is an important priority. The Japanese government's unwillingness to lose face by admitting the extent of the problem in its banking system meant that suspicion and lack of confidence persisted longer than was necessary. When global financial meltdown began in 2008/09, Western governments thought they had learned this lesson, and tried hard to fix the banks. Did they succeed?



RBS share price 2008–12

Source: www.finance.yahoo.co.uk.

Note: In 2012, RBS exchanged 10 old shares for 1 new share. Each 10 shares at 20 pence became 1 share worth 200 pence. There was no change in the value of holdings, merely a 10-fold jump in the

price. The chart above expresses all historical prices as if for new shares, eliminating this artificial jump in the share price in 2012.

By early 2010, many of the banks seemed to be slowly recovering. The chart shows the share price of Royal Bank of Scotland. In 2009, the government took an 84 per cent stake in order to avert its collapse. Its share price, having fallen from 4000 pence to 100 pence, then seemed to stabilize around the 200–300 pence range. The price needs to climb above 500 before the UK government can sell its shares and break even on the original purchase price.

There has been a big difference between the US–UK approach and the approach within the Eurozone. The US was the most aggressive in recapitalizing its banks at the time of the crisis, injecting substantial government funds to restore quickly the solvency of private banks. The UK also undertook major steps to help its banks. In continental Europe, there was generally less injection of public funds to the private banks, leaving some economists worried that Eurozone banks remain fragile and at risk if any new crisis should erupt.

In all countries, improved private sector solvency came at the price of reduced government solvency. Since economies were in recession, government did not raise taxes to pay for the banking bailouts and fiscal stimuli – this would have made recession worse – but rather borrowed money and ran larger budget deficits themselves. Whether governments can cope with high debt burden is a subject to which we return.

This leads to the third lesson from the Japanese experience. If pressing the fiscal accelerator is difficult once the government is heavily indebted, at least the monetary accelerator must be flat on the floor. Japanese monetary policy eventually cut interest rates to zero. Having learned from this experience, central banks in the US, the UK and the Eurozone were quick to slash interest rates to very low levels in 2009 when aggregate demand and output began to plummet. Interest rates have remained low ever since.

The multiplier revisited

The multiplier relates changes in autonomous demand to changes in equilibrium income and output. The formula in Chapter 16 still applies, provided we use c' , the marginal propensity to consume out of gross income, not out of disposable income.

$$\text{Multiplier} = 1/(1 - c')$$

With proportional net taxes t , then $c' = (1 - t)c$, where c remains the marginal propensity to consume out of disposable income, and $(1 - t)$ shows the amount by which disposable income is less than national income. For a given marginal propensity to consume out of disposable income, a higher tax rate t reduces the multiplier. The more the circular flow leaks out into taxation, the less flows round again to stimulate further expansion of output and income. Table 17.1 illustrates.

Table 17.1 Values of the multiplier

C	T	$c' = c(1 - t)$	$\text{Multiplier} = 1/[1 - c'] = 1/[s(1 - t) + t]$
0.9	0	0.90	10.00
0.9	0.2	0.72	3.57
0.7	0	0.70	3.33
0.7	0.2	0.56	2.27
0.7	0.4	0.42	1.72

In Chapter 16, without government the multiplier was simply $1/(1 - c)$ or $1/s$. With a larger marginal propensity to save, there was a larger leakage from the circular flow between firms and households, and the multiplier was correspondingly smaller. Table 17.1 merely extends this insight. Now leakages arise both from saving and from net taxes. When both are large, the multiplier is small. The bottom row of the table has a much smaller multiplier than the top row.

Since $1 - s = c$, Table 17.1 also points out that the denominator of the multiplier $[1 - c(1 - t)]$ is just $s + ct$, which can also be written as $s(1 - t) + t$. From the circular flow, leakages occur through taxation and through saving out of disposable income. The denominator of the multiplier continues to reflect the ‘marginal propensity to leak’, as promised in Chapter 16.

Even Table 17.1 overstates the value of the multiplier in practice. If there was such a large benefit to fiscal expansion, governments would be more willing to boost fiscal policy in order to expand output. By the end of the

chapter, you will understand how leakages from imports further reduce the value of the multiplier.

17.3

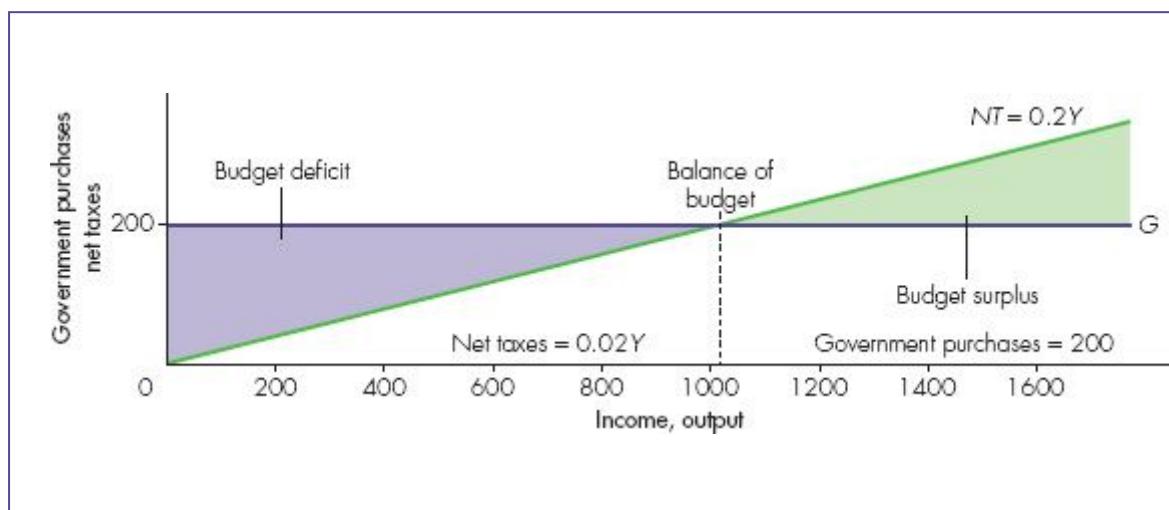
The government budget

The government **budget** describes what goods and services the government will buy during the coming year, what transfer payments it will make and how it will pay for them. Most of its spending is financed by taxes. When spending exceeds taxes, there is a budget deficit. When taxes exceed spending, there is a budget surplus. Continuing to use G for government spending on goods and services, and NT for net taxes or taxes minus transfer payments,

A **budget** is the spending and revenue plan of an individual, a firm or a government.

$$\text{Government budget deficit} = G - NT$$

Figure 17.5 shows government purchases G and net taxes tY in relation to national income. We assume G is fixed at 200. With a proportional net tax rate of 0.2, net taxes are $0.2Y$. At outputs below 1000, the government budget is in deficit; at an output of 1000, the budget is balanced; and at higher outputs, the budget is in surplus.



The budget deficit equals total government spending minus total tax revenue or government purchases of goods and services minus net taxes. Government purchases are shown as constant independent of income, while net taxes are proportional to income. Thus at low levels of income the budget is in deficit and at high income levels the budget is in surplus.

Figure 17.5 The government budget

The budget surplus or deficit is determined by three things: the tax rate t , the level of government spending G and the level of output Y . With a given tax rate, an increase in G will raise output and hence tax revenue. Could the budget deficit be *reduced* by higher spending? We now show that this is impossible.

Investment, saving and the budget

By definition, actual leakages from the circular flow always equal actual injections to the circular flow. Payments cannot vanish into thin air. Our model now has two leakages – saving by households and net taxes paid to the government – and two injections – investment spending by firms and government spending on goods and services. Thus *actual* saving plus *actual* net taxes always equal *actual* government spending plus *actual* investment spending.

In the last chapter we saw that, when the economy is not at equilibrium income, actual saving and investment differ from *desired* or *planned* saving and investment. Firms make unplanned changes in inventories and households may be forced to make unplanned saving if demand exceeds the output actually available.

The economy is in equilibrium when all quantities demanded or *desired* are equal to *actual* quantities. In equilibrium, planned saving S plus planned net taxes NT must equal planned government purchases G plus planned investment I . Planned leakages equal planned injections:

$$S + NT = G + I$$

Without the government, this reduces to the equilibrium condition of Chapter 16: planned saving equals planned investment. Notice that the above equation implies that in equilibrium desired saving minus desired investment equals the government's desired budget deficit:

$$S - I = G - NT$$

A rise in planned government spending G must *raise* the budget deficit. For a given tax rate, a rise in G increases aggregate demand and equilibrium income. Disposable income must rise. Households increase both desired consumption and desired saving.

Since desired investment I is independent of income, this rise in desired saving must increase $(S - I)$ and thus raise $(G - NT)$. This proves that the equilibrium budget deficit rises if government spending increases but the net tax rate is unaltered.

Higher government spending on goods and services increases equilibrium output. With a given tax rate, tax revenue rises but the budget deficit increases (or the budget surplus falls).

We can analyse a tax increase in a similar way. A rise in the tax rate reduces aggregate demand and equilibrium income. Disposable income falls, both because of lower national income and a **higher tax rate**. With less disposable income, desired saving must fall. Since $(S - I)$ is now lower, in equilibrium the budget deficit $(G - NT)$ must also be lower.

For given government spending G , a **higher net tax rate** reduces both equilibrium output and the budget deficit.

We can also understand this more intuitively. When one sector runs a deficit, another sector must be running a surplus to compensate. Saving minus investment is the net surplus of the private sector (households plus firms). A private sector surplus equals a public sector (government) deficit, and vice versa.

17.4 Deficits and the fiscal stance

Is the budget deficit a good measure of the government's **fiscal stance**? Does the size of the deficit show whether fiscal policy is *expansionary*, aiming to raise national income, or *contractionary*, trying to reduce national income?

The **fiscal stance** shows the intended effect of fiscal policy on demand and output.

The deficit may be a poor measure of fiscal stance. The deficit can change for reasons unconnected with a change in fiscal policy. Even if G and t are unaltered, a fall in investment demand reduces output and income, and hence net tax revenue, raising the budget deficit.

For given levels of government spending and tax rates, the budget has larger deficits in recessions, when income is low, than in booms, when income is high. Suppose aggregate demand suddenly falls. The budget will go into deficit. Someone looking at the deficit might conclude that fiscal policy was expansionary and that there was no need to expand fiscal policy further. That might be wrong. The deficit may exist because of the recession.

The structural budget

The **structural budget** shows what the budget will be if output is at potential output.

To indicate the fiscal stance, we calculate the **structural budget**, sometimes known as the *underlying* or *cyclically adjusted* budget. From actual government spending, we subtract not actual net taxes tY but the taxes tY^* that would apply, at the current net tax rate t , if output was hypothetically at potential output Y^* .

The cyclically adjusted budget ($G - tY^*$) is affected by proactive fiscal decisions – changes in government spending G or in net tax rate t , but is insulated from fluctuations in actual tax revenue caused by cyclical deviations of output from potential output. Throughout the recent recession, tax revenue has been disappointingly small. One reason has been that the recovery of output has been slower than hoped and expected. Hence, there has been less tax revenue and more spending on welfare benefits to support citizens in economic distress.

The structural budget depends on the level of potential output Y^* . If a deep recession has permanent effects on potential output itself – for example, firms scrap factories that are then lost forever – the government will have permanently less tax revenue and a permanently larger budget deficit unless other spending is cut or the tax rate increased.

CASE 17.2

CYCCLICAL OUTPUT FLUCTUATIONS AND THE GOVERNMENT BUDGET

The table below shows estimates by the UK Treasury for 2000–15, for fiscal years (April to April). All data are percentages of that period's GDP. The *current budget* includes net tax revenue and government spending on consumption (teachers, nurses, soldiers) but not government investment projects (road building, school building). The total budget deficit, or *net borrowing*, includes government investment expenditure as well as current expenditure on government consumption. By estimating potential output, we can compute cyclically adjusted net tax revenue. The discrepancy between the actual budget and the cyclically adjusted budget is caused only by the [output gap](#).

The [output gap](#) is the percentage deviation of actual output from potential output.

The table shows actual and cyclically adjusted budget deficits of the UK government during 2000–15, both for the total budget and for the current budget – that is, government consumption plus benefit payments less taxes – but excluding government investment. A negative number denotes a budget surplus. It also shows the output gap during the period. A positive number means that output temporarily exceeded potential output; a negative number implies that output was less than potential output.

Budget deficits and the output gap

	Budget deficit [% of GDP]		Cyclically adjusted budget deficit (% of GDP)		Output gap (% of GDP)
	Total	Current	Total	Current	
2000/01	-4	-2.3	1.1	-1.6	1.2
2001/02	0.1	-1.1	0.2	-0.9	-0.1
2002/03	2.4	1.1	1.9	0.6	-0.8
2003/04	2.9	1.5	2.6	1.2	-0.3
2004/05	3.4	1.7	3.1	1.4	-0.3
2005/06	3	1.1	2.8	0.9	-0.2
2006/07	2.5	0.5	2.3	0.4	0.1
2007/08	2.6	0.5	2.6	0.6	0.5
2008/09	6.9	3.6	6.4	3.1	-1.0
2009/10	11.3	7.9	8.9	5.3	-4.2
2010/11	9.6	7	7	4.4	-2.9
2011/12	8	6.2	6.4	4.6	-2.6
2012/13	5.8	6	4	4.2	-2.7
2013/14	5.9	4.5	4.1	2.7	-2.4
2014/15	4.3	3	2.9	1.5	-1.9

The budget deficit rose sharply after the financial crash, peaking in 2009/10. Was this because the recession deprived the government of net tax revenue – less from income tax, VAT and excise duties; more on welfare benefits – or because the government deliberately raised government spending and cut tax rates in order to avert a deeper recession? The table helps us answer this question.

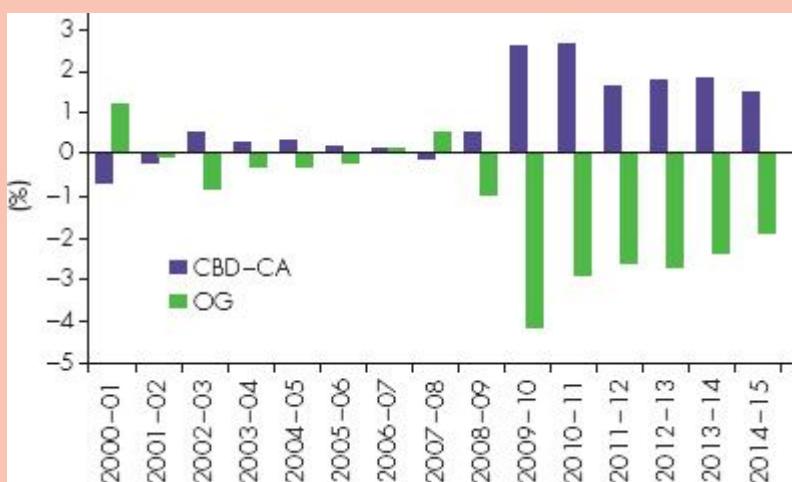
The actual deficit in 2009/10 was 11.3 per cent of GDP, compared with only 2.6 per cent a couple of years earlier. The output gap in 2009/10 was 24.2 per cent, the amount by which actual output was below potential output at the bottom of the recession. The cyclically adjusted total deficit in that year was 8.9 per cent compared with 2.6 per cent a couple of years earlier. What does this imply?

First, the recession deprived the government of net tax revenue equal to 2.4 per cent of GDP – the difference between 11.3 and 8.9 per cent. The recession had a big effect on government finances, but it was only part of the story. Since the cyclically adjusted budget deficit rose from 2.6 per cent to 8.9 per cent in two years, deliberate new policy actions by the government (higher spending, higher subsidies, lower tax rates) accounted for a rise of 6.3 per cent in the government budget. The effect of deliberate fiscal expansion was

more than twice as important as the effect of lower output in causing the actual budget deficit to rise.

The chart below shows more generally the systematic effect of fluctuations in the output gap (shown in green) on the difference between the actual deficit and the cyclically adjusted deficit. This time we focus on the current budget deficit, excluding government investment. The purple columns show the excess of the current budget deficit over its cyclically adjusted measure. The purple columns are about 70 per cent of the height of the green columns. Hence a 1 per cent cyclical fall in output is roughly associated with a 0.7 per cent rise in the current budget deficit (relative to the cyclically adjusted measure), caused by a fall in net tax receipts.

How much the business cycle affects government revenue varies from country to country. For example, at the onset of the financial crash, the US and UK governments both suffered large falls in tax revenue. The fall in Germany was much smaller. The explanation ought to depend on the marginal rate of net taxes – how net tax revenue varies with output and income.



Despite its low-tax rhetoric, the US has a surprisingly ‘progressive’ tax structure that relies heavily on taxing the rich. Germany and most European countries raise most of their revenues through ‘regressive’ consumption and energy taxes – petrol, alcohol, cigarettes, VAT – bearing mainly on the poor and the middle class. Hence, US tax revenue suffers much more in severe recessions, especially if these hit wealthy citizens, such as bankers and stock market investors. Britain’s system lies somewhere in-between, with more reliance on highly redistributive income and capital tax than

Germany but also a much bigger yield than in the US from less progressive taxes on energy and from VAT.

Inflation-adjusted deficits

The **inflation-adjusted budget** uses real not nominal interest rates to calculate government spending on debt interest.

A second reason why actual government deficits may be a poor measure of fiscal stance is the distinction between real and nominal interest rates. The budget deficit treats all nominal interest paid by the government on the national debt as government expenditure on transfer payments. It makes more sense to count only the *real* interest rate multiplied by the outstanding government debt.

Suppose inflation is 10 per cent, nominal interest rates are 12 per cent and real interest rates are 2 per cent. From the government's viewpoint, the interest burden is only really 2 per cent on each £1 of debt outstanding. Although nominal interest rates are 12 per cent, inflation will inflate future nominal tax revenue at 10 per cent a year, providing most of the revenue needed to pay the high nominal interest rates. The real cost of borrowing is only 2 per cent.

Consider what happens when interest rates are 1 per cent (the central bank, worried about recession, has reduced interest rates to a low level) and inflation is 3 per cent (booming China and India keep bidding up the prices of energy and food). The real interest rate is now *minus* 2 per cent. Borrowing now *hurts* the creditor (who gets insufficient interest to cover even inflation) and therefore *helps* the borrower (who by next year will enjoy more than enough inflated tax revenues with which to pay the modest interest charge on the debt).

It is very rare to see estimates of the inflation-adjusted budget. Inflation is therefore governments' secret weapon – it reduces the real burden of the debt unless nominal interest rates exceed the rate of inflation.

17.5 Automatic stabilizers and discretionary fiscal policy

Automatic stabilizers reduce the multiplier and thus output response to demand shocks.

Table 17.1 showed that a higher net tax rate t reduces the multiplier. A high net tax rate is therefore a good automatic stabilizer. Income tax, VAT and unemployment benefit are important **automatic stabilizers**, dampening the output response to changes in autonomous aggregate demand. Automatic stabilizers have a great advantage: nobody has to decide to make any policy decisions. By reducing the responsiveness of the economy to shocks, automatic stabilizers reduce output fluctuations.

All leakages are automatic stabilizers. A higher saving rate (hence a lower marginal propensity to consume) reduces the multiplier. A higher tax rate reduces the multiplier. Later in the chapter, we shall see that a high marginal propensity to import also dampens output fluctuations.

Discretionary fiscal policy is decisions about tax rates and levels of government spending.

Although automatic fiscal stabilizers are always at work, governments also use **discretionary fiscal policies** to change spending levels or tax rates to stabilize aggregate demand. When other components of aggregate demand are abnormally low, the government can boost demand by cutting taxes, raising spending, or both. When other components of aggregate demand are abnormally high, the government raises taxes or cuts spending.

By now you should be asking two questions. First, why can fiscal policy not stabilize aggregate demand completely? Surely, by maintaining aggregate demand at its full-employment level, the government could eliminate booms and slumps altogether? Second, why are governments reluctant to expand fiscal policy and aggregate demand to a level that would completely eliminate unemployment? Concept 17.1 provides some of the answers.

CONCEPT 17.1

THE LIMITS TO FISCAL POLICY

Why can demand shocks not be fully offset by fiscal policy?

1. *Time lags* It takes time to spot that aggregate demand has changed. It may take six months to get reliable statistics on output. Then it takes time to change fiscal policy. Long-term spending plans on hospitals or defence cannot be changed overnight. And once the policy is changed, it takes time to work through the steps of the multiplier process to have its full effect. Where possible, modern economies rely on interest rate changes, not fiscal changes, to make short-term adjustments to aggregate demand.
2. *Uncertainty* The government faces two problems. First, it is unsure of key magnitudes such as the multiplier. It only has estimates from past data. Mistaken estimates induce incorrect decisions about the extent of the fiscal change needed. Second, since fiscal policy takes time to work, the government has to forecast the level that demand will reach by the time fiscal policy has its full effects. If investment is low today but about to rise sharply, a fiscal expansion may not be needed. Mistakes in forecasting nongovernment sources of demand, such as investment, lead to incorrect decisions about the fiscal changes currently required.
3. *Induced effects on autonomous demand* Our model treats investment demand and the autonomous consumption demand as given. This is only a simplification. Changes in fiscal policy may lead to offsetting changes in other components of autonomous demand, as they did in Japan. These induced effects may offset the direct effect of fiscal stimulus if fiscal expansion causes a collapse of confidence because of worries about government debt. If estimates of these induced effects are wrong, fiscal changes have unexpected effects. To study this issue, we extend our model of aggregate demand in Chapter 20.

Why not expand fiscal policy when unemployment is high?

1. *The budget deficit* When output is low and unemployment high, the budget deficit may be large. Fiscal expansion makes it larger. The government may worry about the size of the deficit itself or worry that a large deficit will lead to inflation.
2. *Maybe we are at full employment!* Our simple model assumes there are spare resources. Output is demand-determined. Fiscal

expansion raises demand and output. But we could be at potential output. People are unemployed, and machines idle, only because they do not wish to supply at the going wages or rentals. If so, there are no spare resources to be mopped up raising aggregate demand. If high unemployment and low output reflect not low demand but low supply, fiscal expansion is pointless.

17.6 The national debt and the deficit

Occasionally, governments run a budget surplus. Historically, this is rare. Most governments have budget deficits. The flow of deficits is what adds to the stock of debt. Figure 17.6 shows the history of the UK national debt since the foundation of the Bank of England in 1694.

This figure should be compulsory viewing for all those arguing that today's government debt is 'too high'. For example, the Maastricht Treaty set a ceiling of 60 per cent for the debt/GDP ratio of each member state in the Eurozone. Greece and Japan are thought to be in terrible trouble because their debt GDP ratios are nearly 150 per cent, yet the UK has reached much higher levels on three occasions – 1815, 1918 and 1945. The Brits borrowed in wartime, and paid it off in peacetime.

The 260 per cent debt/GDP ratio of 1815 was the prelude to the biggest boom in British history, nearly a century in which Britain was the undisputed economic powerhouse of the world. Why was Britain not crippled with the burden of the debt?

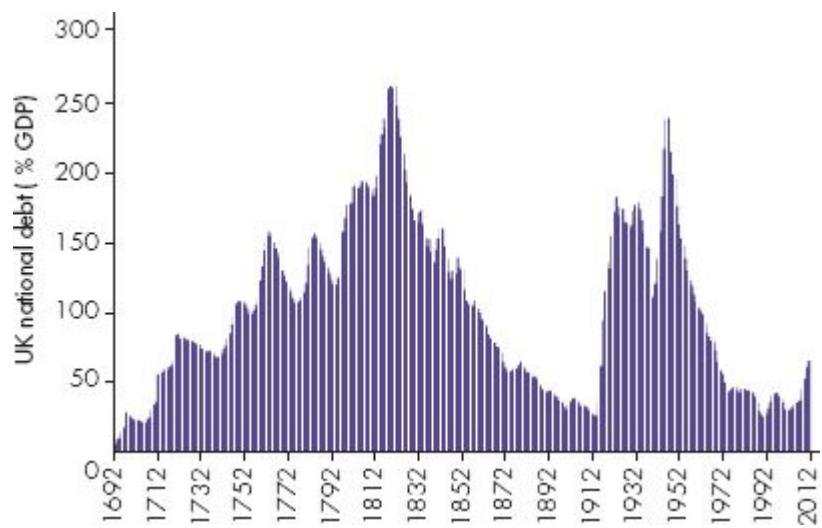


Figure 17.6 UK national debt since 1692 (% of GDP)

Source: en.wikipedia.org/wiki/United_Kingdom_national_debt Available under the Creative Commons Attribution – Share Alike 3.0 Unported License.

We measure debt *relative to GDP* because the latter is a proxy for the likely tax revenue that can be raised at ‘reasonable’ tax rates. Nominal debt rises when there is a budget deficit, but nominal GDP rises because of both real output growth and inflation. Hence, isolating the effect of the inherited debt, the debt burden’s effect on this year’s debt/GDP ratio is given by

$$(r - \pi - g)D/pY$$

where D is nominal debt, p the price level, Y real output, r the nominal interest rate, g the growth rate of real output and π the inflation rate. The debt only hurts if the interest rate r exceeds the rate of nominal income growth ($\pi + g$). The latter shows how quickly the denominator of the debt/GDP ratio is growing; the former shows how interest on inherited debt is increasing the numerator of the debt/GDP ratio.

Since both output growth g and inflation π are normally positive, some steady increase in debt is sustainable without jeopardizing the debt/GDP ratio. Only when interest rates are higher than the rate of nominal income growth does inherited debt create sustainability problems.

This explains what happened to the UK in the nineteenth century. Strong growth and low interest rates made the debt burden negative. The UK

added so much to GDP that the debt/GDP ratio steadily fell. The UK ‘grew its way out of the problem’. Today, Greece, Ireland, Spain and Portugal would love the opportunity to do the same. However, in a global financial market, overseas lenders may panic when they see a country with a high debt/GDP ratio, and therefore raise substantially the interest rate that they charge governments with high debts.

This suggests, correctly, that there may be two possible outcomes, each a self-fulfilling prophecy. First, lenders lose confidence, charge high interest rates, prevent sustainable growth and force the country into a tough choice between a long period of austerity – running budget surpluses despite having low output and low tax revenue. Second, lenders may have faith, keep interest rates low and allow sustainable growth to reduce the debt/GDP ratio. One role of international institutions such as the IMF is to try to engineer the second solution.

Real output growth is one way to solve the problem of an inherited debt burden. A second solution is to print money and create inflation. Rapid inflation benefits the budget in three ways. First, it directly increases nominal GDP and the denominator of the debt/GDP ratio. Future nominal tax revenue increases. Provided interest rates don’t rise by as much as inflation, the real interest rate on the debt is reduced. Second, this applies automatically to that part of the debt that is cash, which has a zero interest rate that cannot rise with higher inflation. So inflation erodes the real value of that part of the debt that is cash rather than bonds. Third, the government cheats. Some taxes are not properly inflation indexed, so that higher inflation actually raises real tax rates. For example, income tax thresholds may not be fully raised in line with inflation, and income tax may apply to nominal interest rates not just to the component that is the real interest rate.

Looking back at Figure 17.6, one reason that the national debt fell in the 1970s, despite little real economic growth, was that double digit inflation eroded the value of nominal debts.

The final way in which a sustainable debt/GDP ratio may be re-established is simply to default on the debt. Creditors hate it, but the day after a large default the government may well be highly solvent and a safer investment from now on than lending to other governments that continue to struggle with high levels of inherited debt. In practice, lenders are likely to be reluctant to lend to a defaulting government for many years thereafter. They are worried the government might default

again, and they wish to send a message to other governments that default is a costly option that will freeze the country out of international financial markets for a long time.

A ‘sustainable’ level of debt depends on the beliefs of lenders, and on their perceptions of the likely behaviour of future governments in the indebted country. What debt level lenders will willingly finance is a matter of politics and psychology as well as economics. In their survey of previous financial crashes and debt crises, Professors Ken Rogoff and Carmen Reinhart conclude that, in practice, the warning bells start sounding when a country’s debt/GDP ratio reaches 90 per cent and/or creditors start charging at least 6 per cent for new loans.³

Figure 17.7 shows the evolution of debt/GDP ratios in a range of countries – the UK, the US, Japan (J), Spain (E) and Germany (D). In 1993, in the green columns, all these countries had debt/GDP ratios below 60 per cent.

Debt was not much of a problem. By 2003, in the orange columns, the big change was the increase in Japan’s government net debt to 78 per cent. After its financial crash in 2003, the government had lost tax revenue due to the subsequent recession, but also had tried to stimulate the economy with a range of discretionary changes to loosen the cyclically adjusted budget.

By 2013, in the purple columns, debt had increased everywhere. Interestingly, in the US it was much higher than in Spain, yet it was Spain that was experiencing the debt crisis. Investors knew, as a last resort, that the UK and the US could create inflation to ease their budgetary problems. Spain, like other Eurozone countries, was bound by the common monetary policy. The European Central Bank could not create inflation for Spain but keep prices stable for Germany.

This completes our introduction to fiscal policy, aggregate demand and the economy. We now extend our model of income determination to include the sector we have so far neglected – foreign trade with the rest of the world.

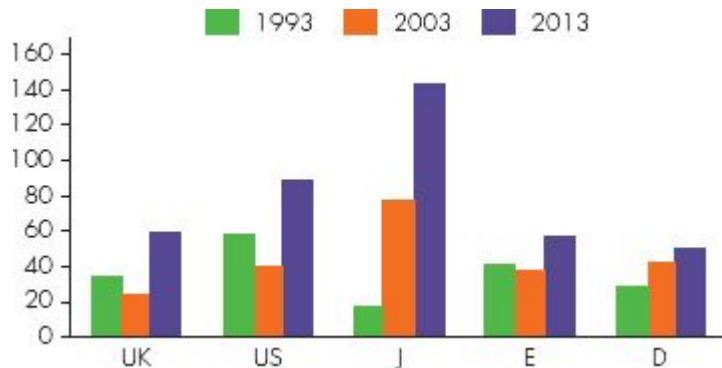


Figure 17.7 Net debt/GDP, 1993–2013 (%)

ACTIVITY 17.1

FISCAL STABILITY, RESPONSIBILITY AND AGGREGATE DEMAND

In 1993, US President Bill Clinton pushed a major tax increase through Congress, yet in subsequent years the US economy recovered from its recession of the early 1990s. Aggregate demand increased. Fiscal conservatives use this as an example of how responsible fiscal policies may actually boost demand, output and employment. We therefore need to explore two issues. First, is this what actually happened in the United States? Second, can it be reconciled with our theoretical model?

The table below shows annual growth of US real GDP, and the level of tax revenue as a percentage of GDP. The US suffered a growth slowdown in 1990 and a small recession in 1991. By 1992 the economy was growing strongly again. With the economy growing, tax revenue should be increasing in absolute terms after 2002. However, with a marginal tax rate of less than 1, tax revenue should grow more slowly than output if tax rates are constant. The *rising* ratio of tax revenue to GDP is indeed evidence of fiscal tightening.

However, the tax rises of 1993 did not initiate output recovery, which began earlier, and it is difficult to know what output growth would subsequently have been if the tax rises had not taken place.

So, empirically, the case for a ‘perverse expansion’ caused by a tighter fiscal policy is at best unproven.

US	1999	2000	2001	2002	2003	2004	2005
Real GDP growth (% per annum)	4.2	1.8	-0.5	3.1	2.7	4.0	2.7
Tax revenue (% of GDP)	29.5	29.3	29.2	28.9	29.2	29.4	29.8

Source: Based on data from Economic Outlook Statistical Annex, © OECD, 2013, <http://www.oecd.org/eco/economicoutlook.htm>, accessed on 18/06/2013.

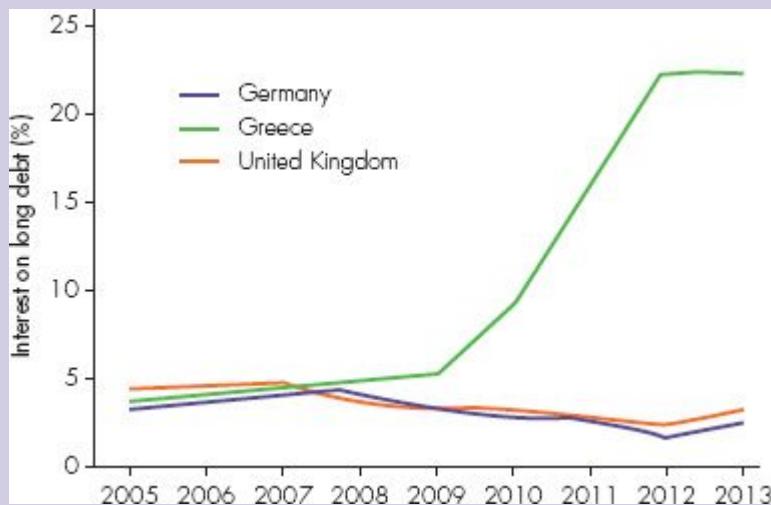
If the example is not empirical proof, what about theoretical arguments. How could a fiscal contraction lead to an output expansion? Implicitly this is a question about which of our model’s assumptions might need to be changed.

Consider the effect on confidence and autonomous aggregate demand. If households and firms believe that the government budget is unsustainable, they worry about a future fiscal crisis. Fearing uncertainty, households cut back on consumption demand and firms postpone investment decisions. In these circumstances, measures to solve the government’s budget problem by increasing taxes or cutting its spending – both of which impact adversely on aggregate demand – could *theoretically* cause a boost to autonomous consumption and investment demand that is large enough to outweigh the contractionary effects of tighter fiscal policy. Usually, this is unlikely. In an acute financial and fiscal crisis, resolving the crisis may indeed generate stability and growth.

A second, related channel is via interest rates. We begin the study of money and interest rates in the next chapter. A country with an unsustainable budget deficit will pay very high interest rates on its debt for two reasons. First, lenders are worried about outright default. Second, the government may deliberately create inflation to erode the real value of debts that are not index-linked. Lenders need to be compensated for both risks.

Very high interest rates reduce consumption demand (households struggle to pay mortgages, house prices may well have collapsed, debt in general is expensive) and investment demand (firms see fewer opportunities to make returns on investment that exceed the high cost of borrowing). Reducing the budget deficit may lead to a sharp reduction in interest rates, boosting autonomous consumption and investment demand enough to offset the contractionary effect of tighter fiscal policy. As with confidence, this channel enters our

simple model by altering the level of autonomous aggregate demand; that is, that part of demand not directly associated with current levels of income and output.



The chart shows long-term interest rates in the UK, Germany and Greece.

The chart makes two points. First, despite being badly affected by the financial crash because of its large financial sector, the UK government enjoyed the same credibility as Germany. Second, investors completely lost confidence in Greece, which faced interest rates of 22 per cent a year, making fiscal policy unsustainable.

This shows the key role played by expectations about the future. If firms, households and market participants believe the future will be satisfactory, they act today in ways more likely to bring that outcome about; if firms and households believe the future will be unsatisfactory, then today they act in order to defend against the bad things they then expect to happen.

How can a government seek to enhance its credibility? It can announce and publicize promises of responsibility that raise the political costs of then backtracking. For example, a *Code for Fiscal Stability* might commit the government to balancing the current budget over the lifetime of the business cycle, borrowing money only for government investment (infrastructure, and so on) that raises future output and future tax revenue. This is sometimes called the *golden rule* of fiscal responsibility, though it is merely

one aspect, since it ignores the effect of inflation and economic growth on the debt/GDP ratio.

Fiscal positions – not just in Mediterranean countries but in countries such as the UK, France and the US – were clearly unsustainable by 2011 after the emergency response to the crash. The issue was and remains *how quickly* the adjustment then has to be made. Eventually, the debt/GDP ratio must stop rising, and ideally be reduced again, which requires that large budget deficits be curtailed. However, too much austerity can be counterproductive if it shrinks tax revenue and makes the budget deficit even harder to close.

Questions

- (a) Why is it important to assess the fiscal position in the medium run and not merely at a point in time?
- (b) If the government can choose the definition of the cycle's length in order to suit its own purposes, why is there a gain from a report by an independent group of fiscal experts in the same way as the Bank of England publishes an independent *Inflation Report* on monetary policy?
- (c) How would you expect financial markets to react if they thought that the government was being misleading in respect of its assessment of whether long-run solvency is gradually being achieved?
- (d) Why does the *golden rule* apply to current expenditure by the government but exclude spending on physical investment?
- (e) Suppose a new government promised to eliminate the large budget deficit within a year to 'put the economy on a sounder footing'. (i) Describe some steps it would have to take on spending and taxes. What effect would this have on: (ii) national output; (iii) tax revenue; (iv) the deficit itself; (v) autonomous investment?

To check your answers to these questions, go to page 679.

17.7 Foreign trade and income determination

Thus far, we have analysed an economy with three sectors – households, firms and the government. We know from our discussion of the national accounts in Chapter 15 that there is a fourth sector, the rest of the world. We now examine output determination once we include the foreign sector.

Exports X are goods and services made at home but sold abroad. Imports Z are goods and services made abroad but bought by domestic residents. Table 17.2 shows UK exports, imports and net exports. Two points should be noted. First, net exports are small relative to GDP, which implies that exports and imports are about equal in size. The UK has fairly balanced trade with the rest of the world. Second, economies are becoming more open as globalization occurs. Both exports and imports have become a larger fraction of GDP during the last 60 years.

Table 17.2 UK foreign trade, 1950–2011 (% of GDP)

	Exports	Imports	Net exports
1950	23	23	0
1970	22	21	1
2011	32	33	-1

Sources: ONS, Economic Trends; www.statistics.gov.uk.

Figure 17.8 confirms that this trend applies across a range of countries. The largest economies, the US and Japan, have a considerable amount of internal trade – enjoying a large domestic market, their producers can specialize and achieve scale economies without the need to export. Large economies are less open to foreign trade. This is not a statement about government policy, merely about the consequences of size itself.

Small economies, such as Ireland and the Netherlands, need to export most of their GDP in order to attain the scale needed to compete in international markets. Their export revenue is used to buy a large quantity of imports. Figure 17.8 shows that imports (and, by implication, exports) account for around 75 per cent of GDP in Ireland and the Netherlands. Larger European countries, such as the UK and France, lie somewhere between these small open economies and the large, quite closed economies. International trade accounts for just over 30 per cent of GDP in France and the UK.

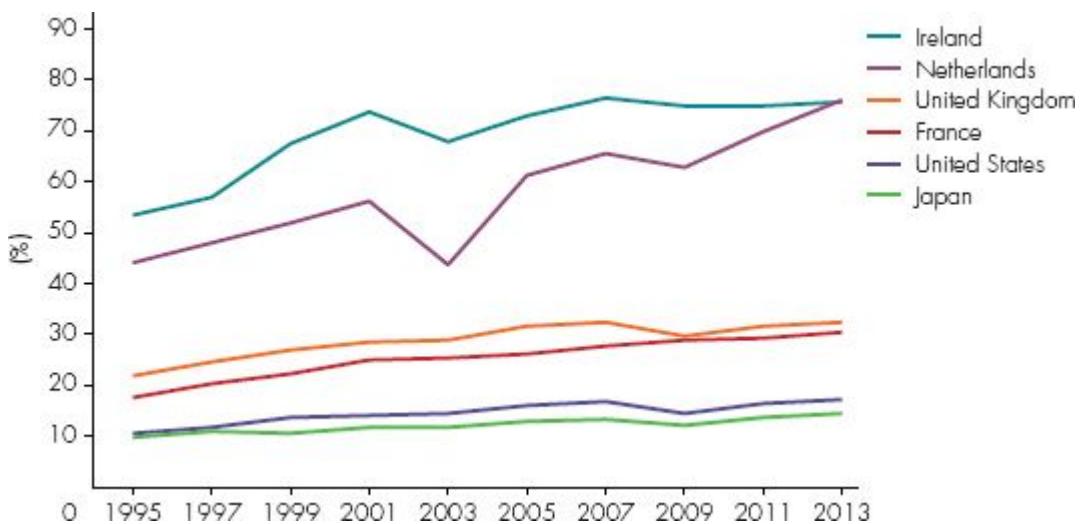


Figure 17.8 The evolution of openness, import/GDP ratios 1995–2013

Source: OECD, Economic Outlook, 2012.

We now explore the consequences of international trade for aggregate demand and output determination. Keep in mind that the importance of the foreign sector for aggregate demand depends a lot on the economy's size, and hence openness.

Net exports $X - Z$ add to income and output. Hence, the equilibrium condition for the goods market must now be expanded to

$$Y = AD = C + I + G + X - Z$$

What determines desired exports and imports? Export demand depends mainly on what is happening abroad. If foreign income and foreign demand are largely unrelated to domestic output, we can treat the demand for exports as autonomous. It does not depend on the level of domestic demand.

The **marginal propensity to import** z is the fraction of each extra pound of national income that domestic residents wish to spend on extra imports.

The **trade balance** is the value of net exports. If this is positive, the economy has a **trade surplus**. If imports exceed exports, the economy

has a **trade deficit**.

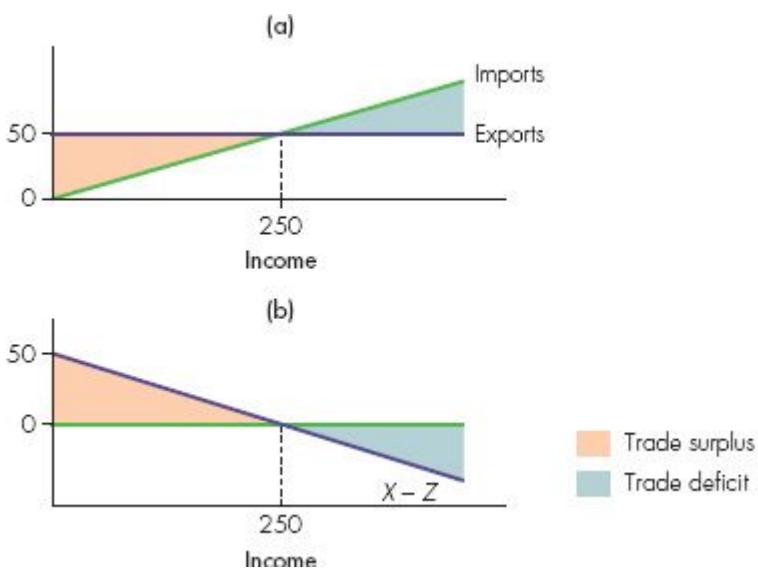
Demand for imports rises when domestic income and output rise. Figure 17.9 shows the demand for exports, imports and net exports, as domestic income changes. The export demand schedule is horizontal. Export demand is independent of domestic income. Desired imports are zero when income is zero but rises as income rises. The slope of the import demand schedule is the **marginal propensity to import**.

The import demand schedule in Figure 17.9 assumes a value of 0.2 for the marginal propensity to import. Each additional pound of national income adds 20 pence to desired imports. This might be true for a very large economy. In small open economies, the marginal propensity to import α is much higher than 0.2. Any increase in national income leads to a large increase in the demand for imports.

At each output, the gap between export demand and import demand is the demand for net exports. At low output, net exports are positive. There is a **trade surplus** with the rest of the world. At high output, there is a **trade deficit** and net exports are negative. By raising import demand while leaving export demand unchanged, higher output worsens the **trade balance**.

Net exports and equilibrium income

Figure 17.10 shows how equilibrium income is determined. We start from the aggregate demand schedule $C + I + G$, described earlier in the chapter, then add net export demand NX , which is simply export demand minus import demand. At low output, net export demand is positive. Aggregate demand $C + I + G + NX$ will then exceed $C + I + G$. As output rises, import demand rises, export demand is constant, so desired net exports fall. At the output of 250, Figure 17.9 told us that net export demand is zero. Figure 17.10 shows the new aggregate demand schedule AD crossing $C + I + G$ at an output of 250. Beyond this output, net export demand is negative and the aggregate demand schedule is below $C + I + G$.

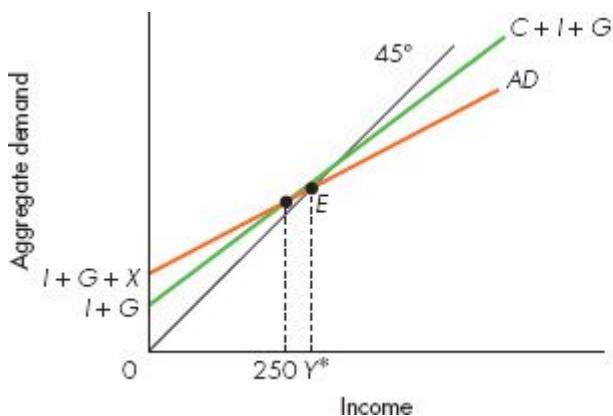


Part (a) shows the given level of exports at 50. Imports increase with the level of income. The diagram assumes a marginal propensity to import, shown by the slope of the import schedule, of 0.2. The trade balance, the difference between planned exports and planned imports, is zero at an income level of 250. Imports and exports both equal 50. At higher levels of income, imports exceed 50 and there is a trade deficit. The net export schedule $X - Z$ in part (b) shows the difference between export and import demand.

Figure 17.9 Exports, imports and the trade balance

At a zero income, Figure 17.10 shows autonomous demand $I + G + X$. Suppose the marginal propensity to consume out of national income MPC is still 0.72. The $C + I + G$ schedule has a slope of 0.72, but the aggregate demand schedule AD is flatter. Each extra pound of national income adds 72 pence to consumption demand but also adds 20 pence to desired imports, since $MPZ = 0.2$. Thus, each extra pound of national income adds only 52 pence to aggregate demand for domestic output. The AD schedule has a slope of 0.52.

In Figure 17.10 equilibrium is at E , where aggregate demand equals domestic income and output. Planned spending, actual incomes and domestic output coincide at Y^* . Knowing this, we can deduce the levels of tax revenue and of imports, and hence compute the budget deficit (or surplus), and the trade deficit (or surplus). Neither is automatically zero merely because the economy is at equilibrium output.



Net exports ($X - Z$) must be added to $(C + I + G)$ to get aggregate demand AD . The gap between $(C + I + G)$ and AD is precisely the net export schedule $(X - Z)$. Equilibrium occurs at E , where the AD schedule crosses the 45° line. Net exports are zero at an income of 250. Thus at Y^* the AD schedule lies below $(C + I + G)$.

Figure 17.10 Equilibrium income in an open economy

The multiplier in an open economy

Each extra unit of national income raises consumption demand for *domestically produced goods* not by c' , the induced additional consumption demand, but only by $(c' - z)$. Some of the extra demand now leaks out into imports without adding to the demand for domestic output. The multiplier is lower because there are leakages not only through saving and taxes but also through imports. In an open economy, the multiplier becomes $1/[1-(c'-z)]$, which is just

$$1/[1 - c' + z]$$

or in full

$$1/[t + s(1-t) + z]$$

As in our previous discussions of the multiplier, it is possible to interpret the denominator as the marginal propensity to leak through all channels – these now occur through taxes, through saving out of disposable income and through imports.

Table 17.3 shows what a difference the inclusion of foreign trade can make to the size of the multiplier. It shows different assumptions about

the number of sectors and different assumptions about the marginal propensities.

Table 17.3 How different parameters affect the multiplier

c	t	$c' = c(1 - t) = t + s(1 - t)$	z	Multiplier $1/[1 - c' + z] = 1/[t + s(1 - t) + z]$	Sectors
0.9	0	0.90	0	10.00	Only firms (F) and households (H)
0.9	0.2	0.72	0	3.57	F, H and government (G)
0.9	0.2	0.72	0.4	1.48	F, H, G and foreign (Fo)
0.7	0	0.70	0	3.33	Only F and H
0.7	0.2	0.56	0	2.27	F, H and G
0.7	0.5	0.35	0.5	0.87	F, H, G and Fo

Without government or foreign sectors, it is easy to get the idea that the multiplier is very large, and aggregate demand very sensitive to changes in autonomous demand. Thus, when $c = 0.9$, the multiplier is 10, and even with $c = 0.7$, the multiplier is 3.33. However, leakages not just through saving but also net taxes reduce the multiplier substantially.

Once we also add imports, and note that the marginal propensity to import can be large in a small open economy, the value of the multiplier is reduced yet further. The final row shows that the multiplier could easily be *less* than 1 (a *divider* rather than a *multiplier*!) if the government and foreign sectors are large enough. The multiplier is still positive – boosts to autonomous demand *do* increase equilibrium output – but by less than you might have supposed. Low multipliers are one reason why proponents of fiscal austerity argue that the output effects will not be too bad.

This matters for economic policy design. A relatively closed economy, such as the US, can engage in fiscal stimulus knowing that (a) the multiplier is a decent size and (b) that only a limited amount of the benefit of the stimulus will leak out into imports. However, in a small open economy, such as Ireland or the Netherlands, most of the effect of fiscal stimulus leaks abroad. Higher domestic autonomous demand still boosts domestic output, but by only a small amount.

Small economies with large governments thus face powerful automatic stabilizers. Whatever shocks occur to autonomous demand are heavily dampened by subsequent leakages to imports and taxation. Larger economies with small governments face fewer leakages and hence, other

things equal, shocks to autonomous demand lead to larger output fluctuations.

Higher export demand

A rise in export demand leads to a parallel upward shift in the aggregate demand schedule AD . Equilibrium income must increase. A higher AD schedule crosses the 45-degree line at a higher level of income. With a higher income, desired imports rise. The analysis of what happens to net exports is very similar to our analysis of the effect of an increase in government spending on the budget deficit.

As a matter of national income accounting, total leakages from the circular flow always equal total injections to the circular flow. And in equilibrium, desired spending must coincide with actual income and spending on domestic goods. Hence the amended equilibrium condition for an open economy is

$$I + G + X = S + NT + Z$$

Desired saving out of disposable income, plus net taxes, plus desired imports, equals desired investment, plus desired government spending, plus desired exports. Higher export demand X raises equilibrium domestic income and output. Because income rises, this raises desired saving, net tax revenue and desired imports. Since S , NT and Z all rise when X rises, the rise in desired imports must be smaller than the rise in desired exports. For example, if export demand rises by 100, desired saving may rise by 20 and net tax revenue by 40, so desired imports can only rise by 40. Net export demand increases. The domestic country's trade balance with the rest of the world improves.

Imports and employment

Do imports steal jobs from the domestic economy? In equilibrium, final demand $C + I + G + X$ is met from domestic output Y and imports Z . By reducing imports, we can create extra output and employment at home. This view is correct, but also dangerous. It is correct because higher consumer spending on domestic rather than foreign goods *will* increase aggregate demand for domestic goods and so raise domestic output and employment.

Figure 17.11 shows planned injections and planned leakages, which are equal in equilibrium at point *A*, at which output is Y . If the government takes action to reduce the propensity to import, shifting the schedule down from Z to Z' , the new equilibrium is at point *B*, at which output is now Y' . With fewer leakages at any output level, it takes higher output and income to generate sufficient planned leakages to equal the unchanged level of planned investment.

There are many ways to restrict import spending at each level of output. In later chapters, we analyse how foreign trade is affected both by the exchange rate and by import tariffs or quotas.

The view that import restrictions help domestic output and employment is dangerous because it ignores the possibility of retaliation by other countries. By reducing our imports, we cut the exports of others. If they retaliate by doing the same thing, the demand for our exports will fall. In the end, nobody gains employment but world trade disappears. If the whole world is in recession, what is needed is a worldwide expansion of fiscal policies, not a collective, and ultimately futile, attempt to steal employment from other countries.

Figure 17.11 can of course also be used to illustrate the effect of a fall in export demand, which we discussed earlier. With lower planned injections, but an unchanged planned leakage schedule, equilibrium output must fall.

If you have been paying close attention, you will have noticed that, in analysing changes in equilibrium output, we sometimes use the diagram with aggregate demand and the 45-degree line, but sometimes use the diagram with planned leakages and planned injections. The two diagrams yield the same answer; which one we use is purely a matter of convenience. Sometimes, the answer is more immediately apparent in one diagram than the other. It is therefore a matter of trial and error which yields the answer more directly and conveniently. However, since it is the same economy being examined, the two diagrams always lead to the same conclusion. If you get a different answer, you made a mistake.

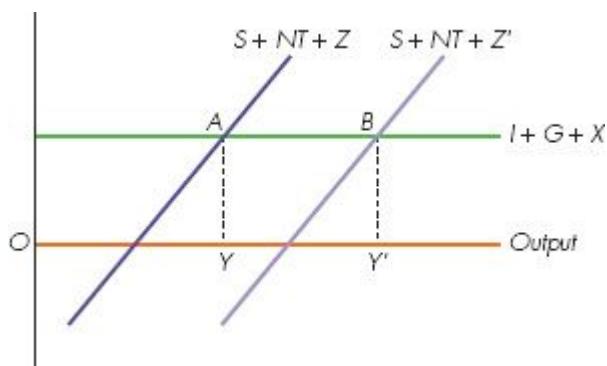


Figure 17.11 A lower import propensity increases equilibrium output

MATHS 17.1

EQUILIBRIUM OUTPUT REVISITED

Short-run equilibrium is given by

$$Y = C + I + G + X - Z$$

where $C = A + c(1 - t)Y$ and $Z = zY$. Hence, $Y = [A + G + X + I] + c(1 - t)Y - zY$, which implies

$$Y = [A + I + G + X] / [1 - c(1 - t) + z] \quad (1)$$

Equilibrium output is the product of autonomous spending – autonomous consumption demand A , plus injections from investment, government spending and exports – and the multiplier $\{1/[1 - c(1 - t) + z]\}$. Because of leakages into saving, taxes and imports, the multiplier may be quite small. In a very small open economy, the marginal propensity to import z will be much higher than in a large closed economy such as the US. Hence the multiplier will be lower in Belgium than in the US. In principle, if the tax rate and marginal propensity to import are large enough, the multiplier could be less than 1. Raising injections by £1 would then raise equilibrium income by less than £1.

Imagine a country such as the US that is large relative to the world economy. It might now need to recognize that US exports depend on how well the world is doing, which in turn is affected by how

much the US imports from the rest of the world, and hence on the level of US output itself. Such interdependence would imply that $X = X^* + fY$, where X^* is autonomous export demand and the positive fraction f measures how much an increase in US output increases exports from the rest of the world, stimulates that economy and thereby increases their import demand for US exports.

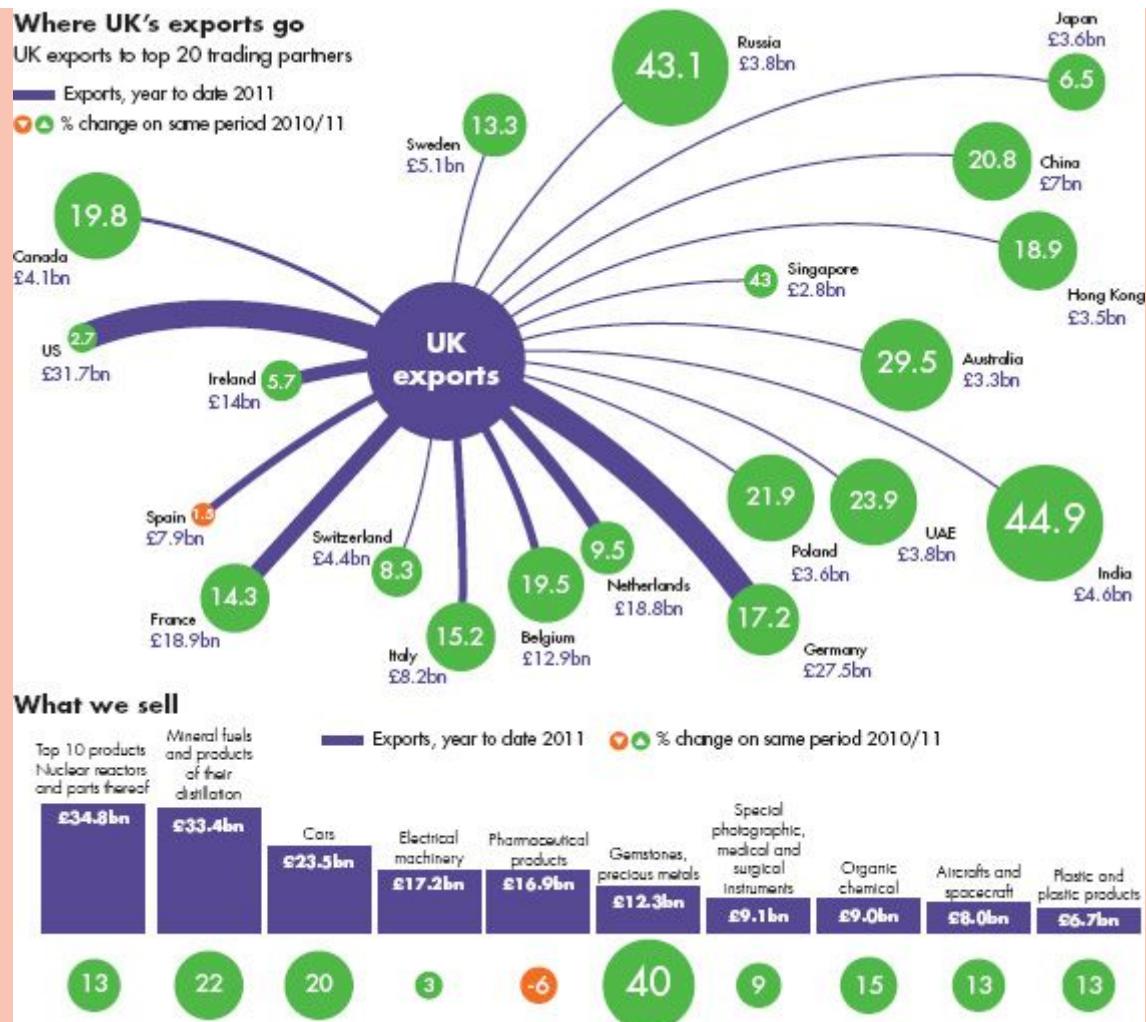
The amended level of equilibrium output, and the product of the corresponding levels of autonomous demand and the new multiplier, would then be

$$Y = [A + I + G + X^*]/[1 - f - c(1 - t) + z]$$

CASE 17.3

GLOBALIZATION AND INTERNATIONAL TRADE

As the global economy becomes more interconnected, to which countries does the UK now export? The figure below provides a snapshot for 2011. The numbers in purple show the value of exports to particular countries and the green (orange) circles show how much higher (lower) this was than in the previous year. £31.7 billion of UK exports still go to the US, but this amount is growing only slowly over time, up just 2.7 per cent from the year before. The Eurozone is now a much more important trading partner, with Germany accounting for £27.5 billion of UK exports, France £18.9 billion and the Netherlands and Belgium together £31.7 billion. When it comes to international trade, countries trade especially strongly with their nearest neighbours. Although trade with emerging market economies is growing much more rapidly, it begins from a much smaller baseline. Hence, in 2011, UK trade with China accounted for only £7 billion of UK exports, and exports to India only £4.6 billion.



Source: © Guardian News & Media Ltd, 2012.

In macroeconomic terms, this means that UK aggregate demand depends most strongly on exports to the Eurozone and hence on how well these countries are doing economically. When the Eurozone expands, it imports more goods from everyone, including the UK; when it contracts, it demands fewer imports and hence UK exports are likely to suffer. Next most important for UK exporting is the economic performance of the US.

UK aggregate demand depends less directly on what is happening in China, but of course there are also indirect effects. If Chinese economic growth slows, not only does this mean fewer Chinese imports from the UK but also fewer imports from other countries such as the US and from the Eurozone. Lower exports to China from the US and the Eurozone lead to lower aggregate demand, as a result of which their demand for imports falls. Hence the UK loses out directly because it exports less to China, but also indirectly because it exports less to the US and the Eurozone. The more

interconnected the world, the harder it is to keep track of all these indirect linkages.

We can think about UK imports in exactly the same way. Data from HM Revenue and Customs show that, in 2011, the UK imported £41 billion worth of goods and services from Germany, £25 billion from the US and £24 billion from China. As with exports, the UK's biggest trading bloc for imports is the Eurozone as a whole. Notice that China features much more strongly in UK imports than it does in UK exports. A UK recession will therefore directly reduce aggregate demand most in the Eurozone, the US and China. The full effect will also depend on all the indirect effects, such as the consequence of lower aggregate demand in the Eurozone then feeding back on its import demand from other countries, which in turn will experience lower demand for their exports.

Summary

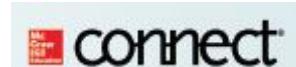
- The government buys goods and services, and levies taxes (net of transfer benefits) that reduce disposable income below national income and output.
- **Net taxes**, if related to income levels, lower the marginal propensity to consume out of national income. Households get only part of each extra pound of national income to use as disposable income.
- **Higher government spending on goods and services** raises aggregate demand and equilibrium output. A **higher tax rate** reduces aggregate demand and equilibrium output.
- An equal initial increase in government spending and taxes raises aggregate demand and output. This is the **balanced budget multiplier**.
- The **government budget** is in deficit (surplus) if spending is larger (smaller) than tax revenue. Higher government spending raises the budget deficit. A higher tax rate reduces it.
- In equilibrium in a closed economy, desired saving and taxes equal desired investment and government spending. An excess of desired

saving over desired investment must be offset by an excess of government purchases over net tax revenue.

- The budget deficit is a poor indicator of **fiscal stance**. Recessions make the budget go into deficit; booms generate a budget surplus. The **structural budget** calculates whether the budget would be in surplus or deficit if output were at potential output. It is also important to **inflation-adjust** the deficit.
- **Automatic stabilizers** reduce fluctuations in GDP by reducing the multiplier. Leakages act as automatic stabilizers.
- The government may also use **active or discretionary fiscal policy** to try to stabilize output. In practice, active fiscal policy cannot stabilize output perfectly.
- Budget deficits add to the **national debt**. If the debt is mainly owed to citizens of the country, interest payments are merely a transfer within the economy. However, the national debt may be a burden if the government is unable or unwilling to raise taxes to meet high interest payments on a large national debt.
- Deficits are not necessarily bad. Particularly in a recession, a move to cut the deficit may lead output further away from potential output. But huge deficits can create a vicious cycle of extra borrowing, extra interest payments and yet more borrowing.
- In an open economy, **exports** are a source of demand for domestic goods but **imports** are a leakage since they are a demand for goods made abroad.
- Exports are determined mainly by conditions abroad and can be viewed as autonomous demand unrelated to domestic income. Imports are assumed to rise with domestic income. The **marginal propensity to import MPZ** tells us the fraction of each extra pound of national income that goes on extra demand for imports.
- Leakages to imports reduce the value of the **multiplier** to $1/[1 - c' + z]$.

- Higher export demand raises domestic output and income. A higher marginal propensity to import reduces domestic output and income.
- The **trade surplus**, exports minus imports, is larger the lower is output. Higher export demand raises the trade surplus; a higher marginal propensity to import reduces it.
- In equilibrium, desired leakages $S + NT + Z$ must equal desired injections $G + I + X$. Thus any surplus $S - I$ desired by the private sector must be offset by the sum of the government deficit ($G - NT$) and the desired trade surplus ($X - Z$).

Review questions



EASY

- 1 In equilibrium, desired saving equals desired investment. Is the statement true or false? Explain.
- 2 Why does the government raise taxes when it could borrow to cover its spending?
- 3 The EU's trading partners are in recession. (a) What happens to the EU's trade balance? (b) What happens to equilibrium EU output? Explain.
- 4 **Common fallacies** Why are these statements wrong? (a) The Chancellor raised taxes and spending by equal amounts. It will be a neutral budget for output. (b) Government policy should balance exports and imports but ensure that the government and private sector spend less than they earn.
- 5 Which of the following statements is correct? The trade surplus equals (a) the government surplus plus the private sector surplus, (b) the government deficit plus the private sector surplus or (c) the government deficit plus the private sector deficit.

MEDIUM

- 6 Equilibrium output in a closed economy is £1000, consumption is £800 and investment is £80. (a) Deduce G . (b) Investment rises by £50. The marginal propensity to consume out of national income is 0.8. What are the new equilibrium levels of Y , C , I and G ? (c) Suppose instead that G

- had risen by £50. What would be the new equilibrium levels of Y , C , I and G ? (d) If potential output is £1200, to what must G rise to make output equal potential output?
- 7 The government spends £6 billion on rail track. The income tax rate is 0.25 and the MPC out of disposable income is 0.8. (a) What is the effect on equilibrium income and output? (b) Assuming that the government budget is in deficit, does the budget deficit rise or fall? Why?
- 8 In 2010, the new UK government wanted to reduce the size of the enormous budget deficit, but also pointed out that the structural budget deficit was significantly smaller than the actual deficit. (a) What does this mean? (b) Why does it matter? (c) Why does this make the subsequent growth of the UK economy so important?
- 9 If Y^{**} is the level of long-run equilibrium output of an open, mixed economy, and if short-run equilibrium output Y^* is given by the model in this chapter, deduce the relationship between the marginal tax rate t , the discrepancy between the actual budget and the structural budget, and the power of the automatic stabilizers.
- 10 Is the ratio of government debt to GDP a useful indicator of a government's indebtedness? When could it be misleading?
- 11 What values of the marginal propensity to save s , the marginal tax rate t and the marginal propensity to import z would be consistent with a multiplier as low as 0.67?

HARD

- 12 **Essay question** ‘By 2007, the UK had had over 50 consecutive quarters of steady growth. This period coincided with the period in which it was decided to make the Bank of England responsible for macroeconomic stabilization. Because interest rates can be changed easily and quickly, whereas tax rates and spending programmes cannot, this example confirms the superiority of monetary policy over fiscal policy in demand management.’ Is this broadly correct? Can you think of examples in which fiscal policy would still be crucial? Did events after 2007 help you answer this question?
- 13 Suppose the marginal propensity to consume out of disposable income is 0.8, the marginal tax rate is 0.5 and the marginal propensity to import is 0.8. Draw a diagram showing the 45-degree line and the aggregate demand schedule. (a) How does this diagram differ from those earlier

in the chapter? (b) What is the size of the multiplier? (c) Illustrate graphically the effect of a shift in aggregate demand.

- |4 Suppose the marginal propensity to consume out of disposable income is 0.8, the marginal tax rate is 0.5 and the marginal propensity to import is 0.8. Draw a diagram showing the 45-degree line and the aggregate demand schedule using the diagram in which planned injections equal planned leakages. (a) How does this diagram differ from those earlier in the chapter? (b) What is the size of the multiplier? (c) Illustrate graphically the effect of a shift in aggregate demand using the diagram in which planned injections equal planned leakages.

- 1 Government is responsible not merely for its own deficits but also for any losses made by state-owned firms. The public sector net cash requirement (PSNCR) is the *government deficit plus net losses of these firms*.
- 2 In contrast, net property income is 1 per cent of GNP. We continue to treat GNP and GDP as equivalent.
- 3 Ken Rogoff and Carmen Reinhart, *This time is different: 8 decades of financial folly* (Princeton University Press, 2011).

CHAPTER 18

Money and banking

Learning outcomes

By the end of this chapter, you should be able to:

- 1 explain the medium of exchange as the key attribute of money
- 2 understand other functions of money
- 3 describe how banks create money
- 4 differentiate between liquidity crisis and solvency crisis
- 5 understand narrow and broad money
- 6 explain the money multiplier and bank deposit multiplier
- 7 recognize different measures of money
- 8 identify motives for holding money
- 9 understand how money demand depends on output, prices and interest rates

Money is a symbol of success, a source of crime and it makes the world go round.

Dogs' teeth in the Admiralty Islands, sea shells in parts of Africa, gold in the nineteenth century: all are examples of money. What matters is not the commodity used but the social convention that it is accepted *without question* as a means of payment. We now explain how society uses money to economize on scarce resources used in the transacting process.

Money is any generally accepted means of payment for delivery of goods or settlement of debt.

18.1 Money and its functions

Although the crucial feature of money is its acceptance as the means of payment or **medium of exchange**, money also has three other functions: a unit of account, a store of value and a standard of deferred payment.

The **medium of exchange** is something accepted as payment only to be subsequently reused to pay for something else.

The medium of exchange

Money is used in almost half of all exchanges. Workers exchange labour services for money. People buy or sell goods for money. We accept money not to consume it directly but to use it subsequently to buy things we do wish to consume. Money is the medium through which people exchange goods and services.¹

To see that society benefits from a medium of exchange, imagine a **barter economy**, in which the seller and the buyer *each* must want something the other has to offer. Each person is simultaneously a seller and a buyer. To see a film, you must swap a good or service that the cinema manager wants. There has to be a *double coincidence of wants*.

A **barter economy** has no medium of exchange. Goods are swapped for other goods.

Trading is very expensive in a barter economy. People spend a lot of time and effort finding others with whom to make mutually satisfactory swaps. Time and effort are scarce resources. A barter economy is wasteful. The use of money – any commodity *generally accepted* in payment for goods, services and debts – makes trading simpler and more efficient. By economizing on time and effort spent in trading, society can use these resources to produce extra goods or leisure, making everyone better off.

Other functions of money

The **unit of account** is the unit in which prices are quoted and accounts kept.

Money is a **store of value** that can be used to make future purchases.

In Britain prices are quoted in pounds sterling; in Germany, in euros. It is convenient to use the same units for the medium of exchange and **unit of account**. However, there are exceptions. During the German hyperinflation of 1922-23, when prices in marks changed very quickly, German shopkeepers found it more convenient to use dollars as the unit of account. Prices were quoted in dollars but payment was made in marks, the German medium of exchange. In 2009 Zimbabwe had to legalize the use of foreign currency as money because its domestic currency was almost worthless after years of hyperinflation.

To be accepted in exchange, money *has to store value*. Nobody will accept money in payment for goods supplied today if the money is worthless when they try to buy goods with it later. But money is not the only, nor necessarily the best, store of value. Houses, stamp collections and interest-bearing bank accounts all serve as stores of

value. Since money pays no interest and its real purchasing power is eroded by inflation, there are better ways to store value.

Finally, money is a *standard of deferred payment* or unit of account over time. When you borrow, the amount to be repaid next year is measured in pounds. However, the key feature of money is its use as a medium of exchange. For this, it must act as a store of value as well. And it is usually, though not invariably, convenient to make money the unit of account and standard of deferred payment as well.

Different kinds of money

In prisoner-of-war camps, cigarettes were money. In the nineteenth century money was mainly gold and silver coins. These are examples of *commodity money*, ordinary goods with industrial uses (gold) and consumption uses (cigarettes), which also serve as a medium of exchange. To use a commodity money, society must either cut back on other uses of that commodity or devote scarce resources to additional production of the commodity. There are cheaper ways for society to make money.

A £10 note is worth far more as money than as a 7.5 X 14 cm piece of high-quality paper. Similarly, the monetary value of most coins exceeds what you would get by melting them down and selling off the metal. By collectively agreeing to use token money, society economizes on the scarce resources required to produce a medium of exchange. Since the manufacturing cost is tiny, why doesn't everyone make £10 notes? The survival of *token money* requires a restriction on the right to supply it. Private production is illegal.² Token money is sometimes called *fiat* money – from the Latin word for 'let it be done' – because its value arises only because of the existence of a government law or regulation.

A *token*(or *fiat*) *money* is a means of payment whose value or purchasing power as money greatly exceeds its cost of production or value in uses other than as money.

An *IOU* *money* is a medium of exchange based on the debt of a private firm or individual.

Society enforces the use of token money by making it *legal tender*. By law, it must be accepted as a means of payment. However, when prices rise very quickly, domestic token money is a poor store of value. People are reluctant to accept it as a medium of exchange. Shops and firms give discounts to people paying in gold or in foreign currency.

In modern economies, token money is supplemented by *IOU* *money*, principally bank deposits, which are debts of private banks. When you have a bank deposit, the bank owes you money. The bank is obliged to pay your cheque. Bank deposits are a medium of exchange because they are generally accepted as payment.

CONCEPT 18.1

BARTER ECONOMY VS MONETARY ECONOMY

Life without money

Some years since, Mademoiselle Zelie, a singer, gave a concert in the Society Islands in exchange for a third part of the receipts. When counted, her share was found to consist of 3 pigs, 23 turkeys, 44 chickens, 5000 cocoa nuts, besides considerable quantities of bananas, lemons and oranges [...] as Mademoiselle could not consume any considerable portion of the receipts herself it became necessary in the meantime to feed the pigs and poultry with the fruit.

(W. S. Jevons, 1898)

This vivid example shows just how costly a barter economy can be. The direct exchange of goods and services for other goods and services either leaves one party with a load of stuff in which they have little interest – in which case they then have to go to the further effort of bartering this in turn for something more useful – or else restricts barter opportunities to the rare cases in which there is a ‘double coincidence of wants’, such that not only does person A want what person B is offering but also person B wants what person A is offering.

The great benefit of a monetary economy is that the medium of exchange can be confidently accepted in the knowledge that it can easily be reused for another transaction. The example below documents the first European to discover paper money. But Europeans did not invent it. As in many other things, the Chinese got there first.

Marco Polo discovers paper money

In this city of Kanbula [Beijing] is the mint of the Great Khan, who may truly be said to possess the secret of the alchemists, as he has the art of producing money. [...]

He causes the bark to be stripped from mulberry trees [...] made into paper [...] cut into pieces of money of different sizes. The act of counterfeiting is punished as a capital offence. This paper currency is circulated in every part of the Great Khan’s domain. All his subjects receive it without hesitation because, wherever their business may call them, they can dispose of it again in the purchase of merchandise they may require.

(*The Travels of Marco Polo*, Book II)

Source: World Bank, *World Development Report*, 1989.

18.2 Modern banking

When you deposit your coat in the theatre cloakroom, you do not expect it to be rented out during the performance. Banks lend out most of the monetary coats in their monetary cloakroom. A theatre would have to get your particular coat back on time, which might be tricky. A bank finds it easier because one piece of money looks just like another.

Unlike other financial institutions, such as pension funds and insurance companies, the key aspect of banks is that some of their liabilities are used as the medium of exchange: cheques allow their deposits to be used as money.

At any time, some people are writing cheques on a Barclays account to pay for goods purchased from a shop that banks with Lloyds; others are writing cheques on Lloyds' accounts to finance purchases from shops banking with Barclays. The *clearing system* is the process of interbank settlement of the net flows required between banks as a result. Thus the system of clearing cheques represents another way in which society reduces the cost of making transactions.³

Liquidity is the cheapness, speed and certainty with which asset values can be converted back into money.

Bank reserves are the money that a bank has available to meet possible withdrawals by depositors.

The money in **sight deposits** can be withdrawn 'on sight' without prior notice.

Time deposits, paying higher interest rates, require the depositor to give notice before withdrawing money.

Private commercial banks have assets and liabilities. Their assets are mainly loans to firms and households, and purchases of financial securities such as bills and bonds issued by governments and firms. Because many securities are very **liquid**, banks can lend short term and still get their money back in time if depositors withdraw their money.

In contrast, many loans to firms and households are quite illiquid. The bank cannot easily get its money back in a hurry. Modern banks thought they could get by with very few **cash reserves** in the vault because they thought they had sufficient liquid assets that would fulfil the same function: in an emergency they could be sold easily, quickly and for a predictable price.

Liabilities of commercial banks include sight and time deposits. Chequing accounts are **sight deposits**. **Time deposits**, which include some savings accounts, pay higher interest rates because banks have time to organize the sale of some of their high-interest assets in order to have the cash available to meet withdrawals. Certificates of deposit (CDs) are large 'wholesale' time deposits – one-off deals with particular clients for a specified period, paying more generous interest rates. The other

liabilities of banks are various 'money market instruments': various types of short-term and highly liquid borrowing by banks.

The business of banking

A bank makes profits by lending and borrowing. To get money in, the bank offers attractive interest rates to depositors, and offers higher interest rates on time deposits than sight deposits since the latter are subject to the possibility of immediate and unpredictable withdrawal.

Banks have to find profitable ways to lend what has been borrowed. In sterling, most is lent as advances of overdrafts to households and firms, usually at high interest rates. Some is used to buy securities, such as long-term government bonds. Some is more prudently invested in liquid assets. Although these pay a lower interest rate, the bank can get its money back quickly if people withdraw a lot of money from their sight deposits. And some money is held as cash, the most liquid asset of all.

A bank uses its specialist expertise to acquire a diversified portfolio of investments. Without the existence of the bank, depositors would have neither the time nor the expertise to decide which of these loans or investments to crisis, modern banks often held reserves as low as 2 per cent of the sight deposits that could be withdrawn at any time. This shows the importance of the other liquid assets in which banks had invested. At very short notice, banks could cash in liquid assets easily and for a predictable amount. The skill in running a bank entails being able to judge how much must be held in liquid assets, including cash, and how much can be lent out in less liquid forms that earn higher interest rates.

Commercial banks are financial intermediaries licensed to make loans and issue deposits, including deposits against which cheques can be written.

A **financial intermediary** specializes in bringing lenders and borrowers together.

A **commercial bank** borrows money from the public, crediting them with a deposit. The deposit is a liability of the bank. It is money owed to depositors. In turn, the bank lends money to firms, households or governments wishing to borrow. Banks are not the only **financial intermediaries**. Insurance companies, pension funds and building societies also take in money in order to re-lend it. The crucial feature of banks is that some of their liabilities are used as a means of payment, and are thus part of the money stock.⁴

18.3 How banks create money

The **reserve ratio** is the ratio of reserves to bank deposits.

To simplify the arithmetic, assume banks use a [reserve ratio](#) of 10 per cent. Suppose, initially, the non-bank private sector has wealth of £1000 held in cash, which is a private sector asset but a liability of the government, who issued it, but not a liability of the private banks. The first row of Table 18.1 shows this cash as an asset of the non-bank private sector.

Table 18.1 Money creation by the banking system

	Banks			N on-bank private sector
	Assets	Liabilities	Monetary assets	Liabilities
Initial	Cash 0 Loans 0	Deposits 0	Cash 1000	Loans from banks
Intermediate	Cash 1000	Deposits 1000	Cash 0 Deposits 1000	Loans from banks
Final	Cash 1000 Loans 9000	Deposits 10 000	Cash 0 Deposits 10 000	Loans from banks 9000

Now people pay this £1000 of cash into the banks by opening bank deposits. Banks have assets of £1000 cash, and liabilities of £1000 of deposits – money owed to depositors. If banks were like cloakrooms, that would be the end of the story. Table 18.1 would end in row 2.

However, banks do not need all deposits to be fully covered by cash reserves. Suppose banks create £9000 of overdrafts. This is a simultaneous loan of £9000, an asset in banks' balance sheets and the granting to customers of £9000 of deposits, against which customers can write cheques. The deposits of £9000 are a liability on banks' balance sheets. Now the banks have £10 000 total deposits – the original £1000 when cash was paid in, plus the new £9000 as counterpart to the overdraft – and £10 000 of total assets, comprising £9000 in loans and £1000 cash in the vaults. The reserve ratio is still 10 per cent in row three of Table 18.1.

It does not even matter whether the 10 per cent reserve ratio is imposed by law or is merely profit-maximizing, smart behaviour by banks that balances risk and reward. The risk is the possibility of being caught short of cash; the reward is the [interest rate spread](#).

How did banks create money? Originally, there was £1000 of cash in circulation. That was the [money supply](#). When paid into bank vaults, it went out of general circulation as the medium of exchange. But the public acquired £1000 of bank deposits against which cheques may be written. The money supply was still £1000. Then banks created overdrafts *not* fully backed by cash reserves. Now the public had £10 000 of deposits against which to write cheques. The money supply rose from £1000 to £10 000. Banks created money.

The [interest rate spread](#) is the excess of the loan interest rate over the deposit interest rate.

The **money supply** is the value of the stock of the medium of exchange in circulation.

ACTIVITY 18.1

A BEGINNER'S GUIDE TO FINANCIAL MARKETS

Financial asset A piece of paper entitling the owner to a specified stream of income for a specified period. Firms and governments raise money by selling financial assets. Buyers work out how much to bid for them by calculating the present value of the promised stream of payments. Assets are frequently retraded before the date at which the original issuer is committed to repurchase the piece of paper for a specified price.

Cash Notes and coins, paying zero interest. It is the most liquid asset.

Bills Short-term financial assets paying no interest directly but with a known date of repurchase by the original borrower at a known price. Consider a three-month UK Treasury bill. In April the government sells a piece of paper, promising to repurchase it for £100 in July. If people bid £98.50 in April, they will make 1.5 per cent in three months by holding the bill until July, when it is worth £100. As July gets nearer, the price at which the bill is retraded climbs towards £100. Buying it from someone else in June for £99.50 and reselling to the government in July for £100 still yields 0.5 per cent in a month, or over 6 per cent a year at compound interest. Treasury bills are easily bought and sold. Their price can only fluctuate over a small range (say, between £98 and £99 in May when they expire in July), so they are highly liquid. People can get their money out easily, cheaply and predictably. Bills issued by companies rather than government are known as *corporate bills*.

Bonds Longer-term financial assets, which again can be issued by companies or by governments as a means of longer-term borrowing. Long-term UK government bonds are known as gilts because their safety is gilt-edged. Look under GILTS – UK CASH MARKET in the second section of the *Financial Times*. You will find a bond listed as Tr 5pc '25, which means that in the year 2025, the UK government will buy back this bond for £100 (the usual repurchase price). Until then, the bondholder gets interest payments of £5 a year (5 per cent of the repurchase price). Bonds are less liquid than bills, not because they are hard to sell, but because the price for which they could be sold, and the cash this will generate, is less certain in the meantime. To see why, we study the most extreme kind of bond.

Perpetuities Bonds never repurchased by the original issuer, who pays interest for ever. These are sometimes called Consols in the UK, because in 1752 the Chancellor and Prime Minister, Sir Henry Pelham, repurchased all the various debts outstanding and financed doing so by issuing one new consolidated bond. Nowadays, 'Consols 2.5%' pay £2.50 a year for ever. Most were issued when interest rates were low, around 2.5 per cent. People originally would have bid around £100 for this Consol. Suppose interest rates on other assets rise to 10 per

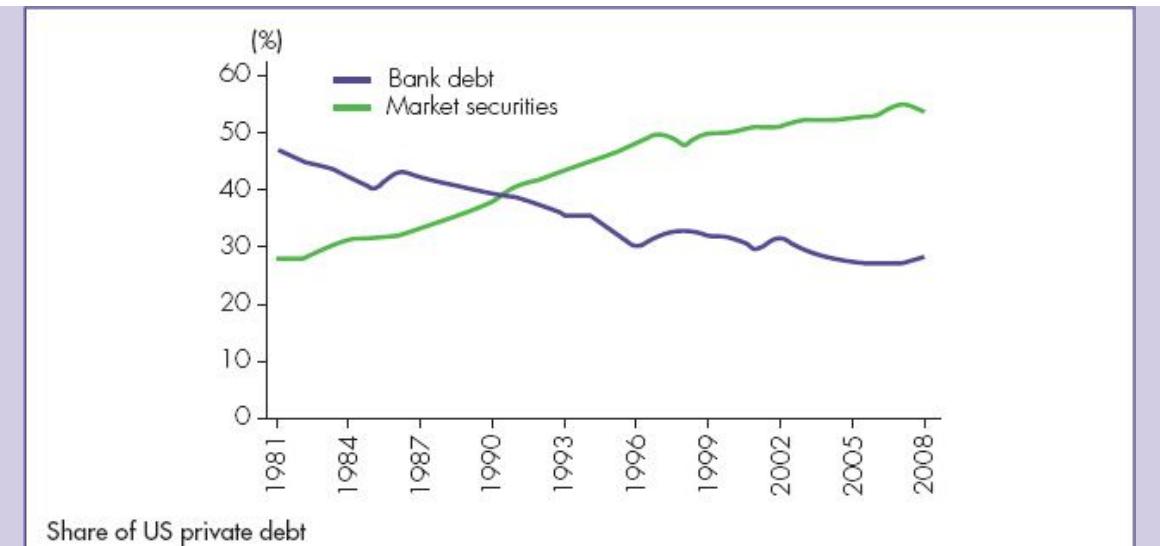
cent. Consols are retraded between people at around £25 each so that new purchasers of these old bonds get about 10 per cent on their financial investment. The person holding a bond makes a capital loss when other interest rates rise and the price of the bond falls. Moreover, since the price of Consols, once £100, could fall to £25 if interest rates rise a lot, Consol prices are much more volatile than the price of Treasury bills. The longer the remaining life of a bond, the more its current price can move around as existing bondholders try to sell on to new buyers at a rate of return in line with other assets today. Bonds can easily be bought and sold, but are not very liquid. You do not know how much you would get if you had to sell out in six months' time.

Company shares (equities) Entitlements to dividends, that part of firms' profits paid out to shareholders rather than retained to buy new machinery and buildings. In good years, dividends are high; in bad years dividends may be zero. Hence a risky asset that is not very liquid. Share prices are volatile. Firms could even go bust, making the shares worthless.

Derivatives Contracts that refer to bets on the value of underlying financial securities. For example, a European aircraft manufacturer might have equities that trade for €300 per share. Somebody might offer to sell a contract paying €50 in three months' time if, but only if, the equity price has risen to at least €350 by that date. The seller of this contract is not very optimistic about the future of the share price. If you think the underlying share price is going to rise beyond €350, you might be happy to buy the derivative contract at a cheap price today. Some derivatives are bets on prices rising; others are bets on price falling.

Securitization The practice of aggregating collections of individual contracts (such as individual mortgages) into bundles of contracts that are then sold and bought by institutions far removed from the original deal. In theory, securitization was meant to spread the risk of an individual contract going wrong, thereby reducing risk in total. In practice, many of the buyers of these securities understood little about them and were amazed when the contracts all became worthless together. The most famous example was the securitization of US sub-prime mortgages – dubious loans to poor people who were often duped into taking out mortgages whose repayments they would later be unable to afford. Suddenly, many institutions around the world found themselves holding 'assets' that were revealed to be worth almost nothing, setting up a tsunami of insolvency.

The chart below shows the huge growth of securitization, and corresponding fall in old-fashioned bank loans, in the US in the run-up to the financial crash.



Source: US Federal Reserve.

Questions

- (a) If cash pays no interest, why does anyone hold it?
- (b) Since firms could use bills and bonds to raise finance, what advantages do they see in raising money through issuing equities?
- (c) If it is good for firms to issue equities, can it simultaneously be good for investors?
- (d) While the Treasury is closed and the prime minister is on holiday, the Bank of England announces it has made a loan to a regional bank whose depositors were panicking. Did the Bank of England think this was a liquidity crisis or a solvency crisis? Explain your answer.

To check your answers to these questions, go to page 679.

18.4

The traditional theory of money supply

We begin the traditional account of the role of banks in the money supply. The *central bank* – the Bank of England in the UK, the ECB in the Eurozone, the Federal Reserve in the US – controls the issue of token money in a modern economy. Private creation of token money must be outlawed when its value as a medium of exchange exceeds the direct cost of its production.

People hold cash for many reasons. It makes transactions easier and cheaper. Moreover, some people do not trust banks; they keep their savings under the bed. Remarkably, only three-quarters of British households have chequing accounts. Some people hold cash in order to make illegal or tax-evading transactions in the ‘black economy’. In a modern economy most of the broad measure of money is in bank deposits.

Suppose, as was the case a few years ago, UK banks hold cash reserves equal to 1 per cent of their total deposits, and the private sector holds cash in circulation equal

to 3 per cent of the value of sight deposits. Maths 18.1 shows that this implies a **money multiplier** of 26. Each £100 rise in the monetary base increases the money supply by £2600.

The **money multiplier** is the ratio of broad money to the monetary base.

MATHS 18.1

THE MONEY MULTIPLIER

Suppose banks wish to hold cash reserves R equal to some fraction of deposits D , and that the private sector holds cash in circulation C equal to a fraction C_p of deposits D :

$$R = c_b D \text{ and } C = c_p D$$

The monetary base H is either in circulation or in bank vaults:

$$H = C + R = (c_p + c_b)D$$

Finally, the money supply is circulating currency C plus deposits D :

$$M = C + D = (c_p + 1)D$$

These last two equations give us the money multiplier, the ratio of M to H :

$$M/H = (c_p + 1)/(c_p + c_b) > 1$$

If the public hold cash to the value of 2 per cent of their deposits, $c_p = 0.02$, and if banks hold reserves equal to 2 per cent of deposits, $c_b = 0.02$. Hence the money multiplier is

$$M/H = 1.02/(0.04) = 25.5$$

The **bank deposit multiplier** is the ratio of broad money to bank reserves.

the ratio of broad money to the monetary base. The ratio M/R , dividing broad money only by bank reserves but not by cash held outside the banks, is called the **bank deposit multiplier**:

$$M/R = (c_p + 1)/(c_b) = 1.02/0.02 = 51$$

Notice that, if banks now become cautious about lending and raise their reserve ratio c_b to 10 per cent of deposits, then

$$\text{Money multiplier} = M/H = 1.02/0.12 = 8.5$$

$$\text{Bank deposit multiplier} = M/R = 1.02/0.10 = 10.2$$

so the values of both the money multiplier and bank deposit multiplier are considerably smaller. If the government wishes to maintain the previous level of broad money M , it will need to inject considerably more narrow money than previously.

The **bank deposit multiplier** is the ratio of broad money to bank reserves

For now, it is important to remember that a fall in either the banks' desired cash reserve ratio or the private sector's desired ratio of cash to bank deposits raises the money multiplier (and the bank deposit multiplier). For a given monetary base, the money supply rises.

What determines the cash reserve ratio desired by banks? The higher the interest rate spread, the more banks wish to lend and the more they risk a low ratio of cash reserves to deposits. Conversely, the more unpredictable are withdrawals from deposits, or the fewer lending opportunities banks have in very liquid loans, the higher cash reserves they have to maintain for any level of deposits.

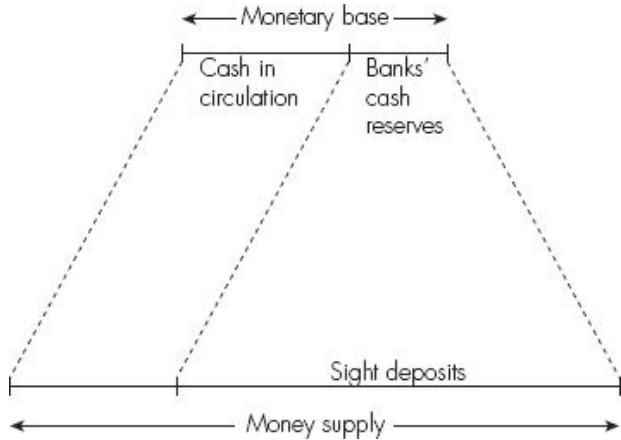
The public's desired ratio of cash to deposits partly reflects institutional factors, for example whether firms pay wages by cheque or by cash. It also depends on the incentive to hold cash to make untraceable payments to evade taxes. And credit cards reduce the use of cash. Credit cards are a temporary means of payment; that is, a *money substitute* not money itself. A signed credit card slip cannot be used for *further* purchases. Soon, you have to settle your account using money. Nevertheless, since credit cards allow people to carry less cash in their pocket, their increasing use reduces the desired ratio of cash to bank deposits.

Figure 18.1 summarizes the traditional account of the **monetary base** and the money supply. The monetary base, or stock of high-powered money, is held either as cash reserves by the banks or as cash in circulation. Since bank deposits are a multiple of banks' cash reserves, the money multiplier exceeds unity. The money multiplier is larger (a) the lower the non-bank public's desired ratio of cash to bank deposits, giving the banks more cash with which to create a multiplied deposit expansion, and (b) the lower is the banks' desired ratio of cash to deposits, leading them to create more deposits for any given cash reserves.

The **monetary base**, or **narrow money**, is the quantity of notes and coins in private circulation plus the quantity of reserves held by commercial banks.

How does the central bank change the level of the monetary base? It is the monopoly supplier of cash, but it does not get it into circulation by dropping it from a

helicopter. Rather, it buys financial assets in exchange for cash that it has itself created, or it sells financial assets in exchange for cash which it then retires from use. We discuss this process more fully in the next chapter.



The money supply comprises currency in circulation and deposits at banks. The monetary base, issued by the central bank, is held either as currency in circulation or as banks' cash reserves. Since deposits are a multiple of banks' cash reserves, the money multiplier exceeds 1. The monetary base is 'high-powered' because part of it is multiplied up as the banking system creates additional deposits, the major component of the money supply.

Figure 18.1 Traditional money supply determination

CONCEPT 18.2

THE COLLAPSE OF BANK LENDING

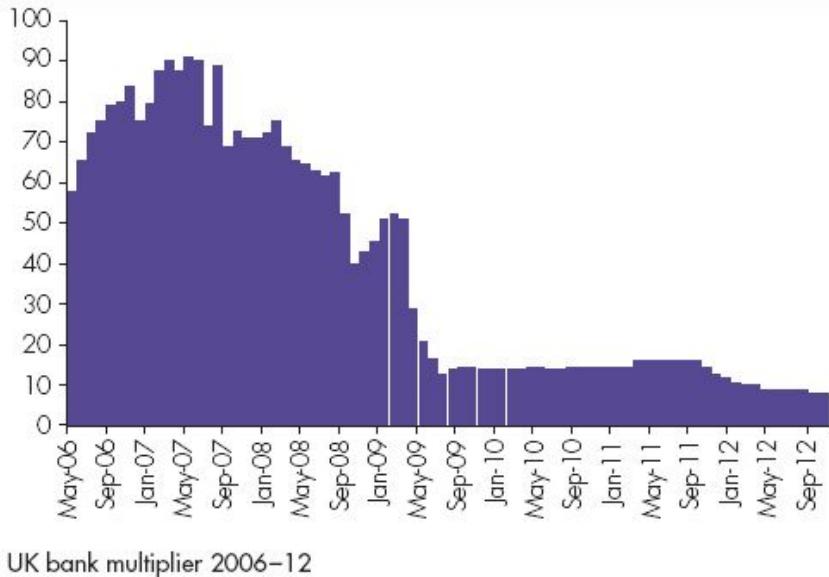
The bank deposit multiplier

The money multiplier is $M/(R + C_p)$, the broad money supply M divided by banks' reserves R and cash held by the public C_p . Prior to the financial crisis, it had a value of around 26. The bank deposit multiplier M/R is several times larger, since the denominator is smaller. The figure shows that, just before the crisis, the bank deposit multiplier was around 90. Banks had £90 of deposits for every £1 in reserves.

However, the more bank lending dried up during the credit crunch, the smaller the bank deposit multiplier became. For any given level of reserves, **broad money** was much smaller. The figure shows that the bank deposit multiplier fell from 90 to around 14 in 2010 and to below 10 by 2012.

This has two implications. First, the theory of monetary control, which we discuss in Chapter 19, cannot assume that financial multipliers are stable over time: they are very sensitive to the level of confidence in banks and the extent of risk-taking they feel able to undertake.

Broad money includes all assets fulfilling the functions of money, and is principally bank deposits.



Second, without some additional change in behaviour by the Bank of England, the collapse of the bank deposit multiplier threatened to lead to a corresponding collapse of broad money. The attempt to prevent this, known as quantitative easing, is discussed in Chapter 19.

Measures of money

Money is the medium of exchange available to make transactions. Hence, the money supply is cash in circulation outside banks, plus bank deposits. It sounds simple, but is not. Two issues arise: which bank deposits, and why only bank deposits?

We can think of a spectrum of liquidity. Cash, by definition, is completely liquid. Sight deposits (chequing accounts) are almost as liquid. Time deposits (savings accounts) used to be less liquid, but now many banks offer automatic transfer between savings and chequing accounts when the latter run low. Savings deposits are almost as liquid as chequing accounts.

UK statistics distinguish between *retail* and *wholesale* deposits. Retail deposits are made in high-street branches at the advertised rate of interest. Wholesale deposits, big one-off deals between a corporate depositor and a bank at a negotiated interest rate, are also quite liquid.

Everyone used to be clear about what a bank was, and hence whose deposits counted towards the money supply. Financial deregulation blurred this distinction in the UK and the US, and is now doing so in continental Europe. Before 1980, UK banks did not lend for house purchase, and cheques on building society deposits could not be

used at the supermarket checkout. Now ‘banks’ compete vigorously for mortgages, supermarket chains are in the banking business and building society cheques are widely accepted as a means of payment. There is no longer a reason to exclude building society deposits from measures of the money supply. Since January 2010, UK monetary statistics do not even distinguish between banks and building societies.

Table 18.2 shows the components of broad money in the UK in 2009. Notes and coins in circulation outside the Bank of England are the most liquid form of the medium of exchange. To this we add retail deposits in banks and building societies. Next, we add wholesale deposits. The sum of all these is M4, the most commonly used measure of broad money.

Table 18.2 Narrow and broad money in the UK, October 2012

	£ billion
cash in circulation (outside central bank)	46
+ retail deposits in banks and building societies	1134
+ wholesale deposits	867
= Money supply M4 (broad money)	2047

Source: Bank of England.

18.5 The demand for money

In most advanced countries, the quantity of broad money is nearly 100 times larger than in 1960. Why do we hold so much extra money? We focus on three variables that affect money demand: interest rates, the price level and real income.

The **demand for money** is a demand for *real* money balances.

Motives for holding money

Money is a stock. It is the quantity of circulating currency and deposits *held* at any given time. Holding money is not the same as *spending* it. We hold money now to spend it later.

Money is the medium of exchange, for which it must also be a store of value. These two functions of money provide the reasons why people wish to hold it. People can hold their wealth in various forms – money, bills, bonds, equities and property. For simplicity, assume that there are only two assets: money, the medium of exchange that pays no interest, and bonds, which we use to stand for all other interest-bearing assets that are not directly a means of payment. As people earn income, they add to their wealth. As they spend, they deplete their wealth. How should people divide their wealth between money and bonds?

People **hold money** only if there is a benefit to offset this cost. What is that benefit?

The **cost of holding money** is the interest given up by holding money rather than bonds.

The transactions motive

Transacting by barter is costly in time and effort. Holding money economizes on these costs. If all transactions were perfectly synchronized, we could be paid at the same instant as we did our spending. Except at that instant, we need hold no money at all.

Must we hold money between being paid and making subsequent purchases? We could put our income into interest-earning assets, to be resold later when we need money for purchases. However, every time we buy and sell assets there are brokerage and bank charges. And it takes an eagle eye to keep track of cash flow and judge the precise moment at which money is needed and assets must be sold. If small sums are involved, the extra interest does not compensate for the brokerage fees, and the time and effort. It is easier to hold some money.

The **transactions motive** for holding money reflects the fact that payments and receipts are *not* synchronized.

How much money we need to hold depends on the value of the transactions we later wish to make and the degree of synchronization of our payments and receipts. Money is a nominal variable not a real variable. How much £100 buys depends on the price of goods. If all prices double, our receipts and our payments double in nominal terms. To transact as before we need to hold twice as much money.

We need a given amount of real money, nominal money deflated by the price level, to make a given quantity of transactions. When the price level doubles, other things equal, the demand for nominal money balances doubles, leaving the demand for real money balances unaltered. People want money because of its purchasing power in terms of the goods it will buy.

Real GNP is a good proxy for the total real value of transactions. Thus we assume that the transactions motive for holding real money balances rises with real income.

The transactions motive for holding money also depends on the synchronization of payments and receipts. Suppose, instead of shopping throughout the week, households shop only on the day they get paid. Over the week, national income and total transactions are unaltered, but people now *hold* less money over the week.⁵

A nation's habits for making payments usually change only slowly. In our simplified model we assume that the degree of synchronization is constant over time. Thus we focus on real income as *the* measure of the transactions motive for holding *real* money balances.

Of course, the degree of synchronization is not literally constant over time. For example, compared with having to queue up in a bank to withdraw cash from one's

account, the introduction of ATMs (cash machines) made it easier to get cash, even when banks were closed. People therefore held less cash on average in their pocket because they could easily get more. Conversely, bank accounts were even more useful than previously.

The precautionary motive

We live in an uncertain world. Uncertainty about the timing of receipts and payments creates a precautionary motive for holding money.

Suppose you buy a lot of interest-earning bonds and get by with a small amount of money. Walking down the street you see a great bargain in a shop window, but have too little money to close the deal. By the time you cash in some bonds, the bargain is gone, snapped up by someone with ready money.

How can we measure the benefits from holding money for precautionary reasons? The payoff grows with the volume of transactions we undertake and with the degree of uncertainty. If uncertainty is roughly constant over time, the level of transactions determines the benefit of real money held for precautionary reasons. As with the transactions motive, we use real GNP to proxy the level of transactions. Thus, other things equal, the higher is real income, the stronger is the **precautionary motive** for holding money.

In an uncertain world, there is a **precautionary motive** to hold money. In advance, we decide to hold money to meet contingencies that we cannot yet foresee.

The transactions and precautionary motives are the main reasons to hold the medium of exchange, and are most relevant to the benefits from holding a narrow measure of money. The wider measure, M4, includes higher-interest-earning deposits. The wider the definition of money, the less important are the transactions and precautionary motives that relate to money as a medium of exchange, and the more we must take account of money as a store of value.

The asset motive

Forget the need to transact. Think of someone deciding in which assets to hold wealth. At some distant date, wealth may be spent. In the short run, the aim is a good but safe rate of return.

Some assets, such as company shares, on average pay a high return but are risky. Some years their return is *very* high, in other years it is negative. When share prices fall, shareholders make a capital loss that swamps the dividends they receive. Other assets are less risky, but their average rate of return is correspondingly lower.

The **asset motive** for holding money reflects dislike of risk. People sacrifice a high average rate of return to obtain a portfolio with a lower but safer return.

How should people divide their portfolios between safe and risky assets? You might like to reread Chapter 13. Since people dislike risk, they will not put all their eggs in one basket. As well as holding some risky assets, they will keep some of their wealth in safe assets.

The **asset motive** for holding money is important when we consider why people hold broad measures of money such as M4.

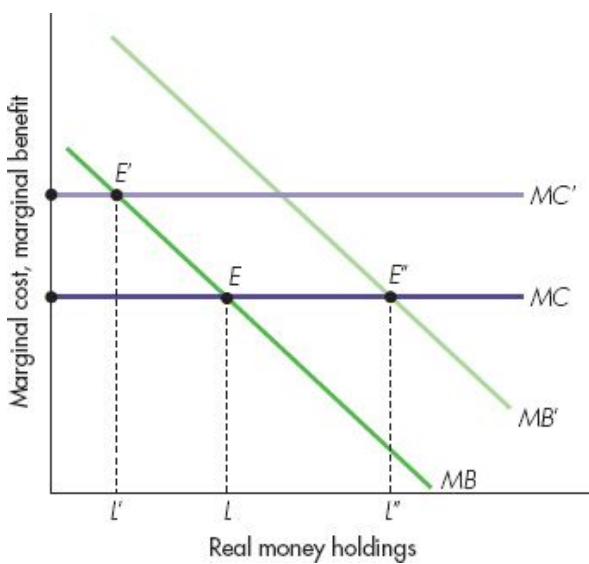
The demand for money: prices, real income and interest rates

The transactions, precautionary and asset motives suggest that there are benefits to holding money. But there is also a cost, the interest forgone by not holding high-interest-earning assets instead. People hold money up to the point at which the marginal benefit of holding another pound just equals its marginal cost. Figure 18.2 illustrates how much money people want to hold.

People want money for its purchasing power over goods. The horizontal axis plots real money holdings; that is, nominal money in current pounds divided by the average price of goods and services. The horizontal line MC is the marginal cost of holding money; that is, the interest forgone by not holding bonds. MC shifts up if interest rates rise.

The MB schedule is the marginal benefit of holding money. We draw MB for a given real GDP measuring the transactions undertaken. For this level of transactions, it is possible but difficult to get by with low real money holdings. We have to watch purchases and receipts and be quick to invest money as it comes in and ready to sell off bonds just before we make a purchase. Nor do we have much precautionary money. We may be frustrated or inconvenienced if, unexpectedly, we want to make a purchase or settle a debt.

With low real money holdings, the marginal benefit of another pound is high. We can put less effort into timing our transfers between money and bonds, and have more money for unforeseen contingencies. For a given real income and level of transactions, the marginal benefit of the last pound of money holdings declines as we hold more real money. With more real money, we have plenty both for precautionary purposes and for transaction purposes. Life is easier. The marginal benefit of yet more money holding is low.



The horizontal axis shows the purchasing power of money in terms of goods. The MC schedule shows the interest sacrificed by putting the last pound into money rather than bonds. The MB schedule is drawn for a given real income and shows the marginal benefits of the last pound of money. The marginal benefit falls as money holdings increase. The desired point is E , at which marginal cost and marginal benefit are equal. An increase in interest rates, a rise in the opportunity cost schedule from MC to MC' , reduces desired money holdings from L to L' . An increase in real income increases the marginal benefit of adding to real balances. The MB schedule shifts up to MB' . Facing the schedule MC , a shift from MB to MB' increases real money holdings to L'' .

Figure 18.2 Desired money holdings

Given our real income and transactions, desired money holdings are at E in Figure 18.2. For any level of real money below L , the marginal benefit of another pound exceeds its marginal cost in interest forgone. We should hold more money. Above L , the marginal cost exceeds the marginal benefit and we should hold less. The optimal level of money holding is L .

To emphasize the effect of prices, real income and interest rates on the quantity of money demanded, we now change each of these variables in turn. If all prices of goods and services double but interest rates and real income are unaltered, neither MC nor MB shifts. The desired point remains E and the desired level of *real* money remains L . Since prices have doubled, people hold twice as much nominal money to preserve their real money balances at L .

If interest rates on bonds rise, the cost of holding money rises. Figure 18.2 shows this upward shift from MC to MC' . The desired point is now E' and the desired real money holding falls from L to L' . Higher interest rates reduce the quantity of real money demanded.⁶

Finally, consider a rise in real income. At each level of real money holdings, the marginal benefit of the last pound is higher than before. With more transactions to undertake and a greater need for precautionary balances, a given quantity of real money does not make life as easy as it did when transactions and real income were lower. The benefit of a bit more money is now greater. Hence, when real income

rises, in Figure 18.2 we can show this as an upward shift in marginal benefit from MB to MB' .

At the original interest rate and MC schedule, the desired level of money balances is L_0 . Thus a rise in real income raises the quantity of real money balances demanded. Table 18.3 summarizes our discussion of the demand for money as a medium of exchange.

Table 18.3 The demand for money

Quantity demanded	Effect of rise in		
	Price level	Real income	Interest rate
Nominal money	Rises in proportion	Rises	Falls
Real money	Unaffected	Rises	Falls

So far we have studied the demand for cash, the narrowest measure of money. Wider definitions of money must also recognize the asset motive for holding money. To explain the demand for M4, we interpret MC as the average extra return by putting the last pound into risky assets rather than time deposits, which are safe but yield a lower return. For a given wealth, MB is the marginal benefit of time deposits in reducing the risk of the portfolio. If no wealth is invested in time deposits, the portfolio is very risky. A bad year is a disaster. There is a big benefit in having some time deposits. As the quantity of time deposits increases, the danger of a disaster recedes and the marginal benefit of more time deposits falls.

A rise in the average interest differential between risky assets and time deposits shifts the cost of holding broad money from MC to MC' , reducing the quantity of broad money demanded. Higher wealth shifts the marginal benefit from MB to MB' . More time deposits are demanded.

Explaining the rise in UK money holdings from 1965 to 2012

Why were nominal money holdings 90 times higher in 2012 than in 1965? We have identified three explanations: prices, real income and nominal interest rates. Table 18.4 shows how these variables changed

over the period.

Table 18.4 Holdings of M4, 1965–2012

	1965		2012
<i>Index of:</i>			
Nominal M4	100	9494	
Real M4	100	720	
Real GDP	100	290	
Interest rate (%)	6	1	

Although nominal money holdings rose 95-fold, the price level also rose a lot between 1965 and 2012. Table 18.4 shows real money rising more than seven-fold over the period. Real GDP was almost three times its initial level. Higher real output and income raised the quantity of real money demanded. Nominal interest rates fell substantially, which also added to the demand for money.

To sum up, most of the increase in desired holdings of nominal money was merely because prices were higher: it took more nominal money merely to maintain the purchasing power of money holdings. Real economic growth caused an increase in real GDP and in real income: this led to a significant increase in the demand for real money, but cannot explain the whole of that increase. The remaining part is explained by a lower opportunity cost of holding money: as interest rates fell, it was no longer so expensive to allocate wealth to money rather than to other interest-bearing assets.

18.6 Financial crises

A **financial panic** is a self-fulfilling prophecy. Believing a bank will be unable to pay, people rush to get their money out. But this makes the bank go bankrupt.

Everybody knows what the banks are doing. Usually, people do not mind. But if people believe that a bank has lent too much and will be unable to meet depositors' claims, there will be a *run* on the bank – a **financial panic**. If the bank cannot repay all depositors, you try to get your money out first while the bank can still pay. Since everyone does the same thing, they ensure that the bank is unable to pay. Some of its loans will be too illiquid to get back in time.

Notice that there are two kinds of financial crisis. First, a bank may have made loans that turn out to be worthless. They are no longer valuable assets of the bank. Liabilities now exceed assets and the bank is insolvent. Unless rapidly bailed out by injections of new assets by shareholders or the government, the bank will be declared bankrupt and it will be closed down. In such circumstances, depositors are simply being smart in trying to get their money out before this happens.

However, there may also be self-fulfilling panics even when the bank's assets are fine and the bank is not insolvent. If a depositor believes that other depositors will panic and withdraw money, it makes no sense to be last in the queue – the bank may have trouble selling enough liquid assets quickly enough to meet all the withdrawals, and it may be forced into difficulty by the panic itself.

We call this second case a **crisis of liquidity**, whereas the first case is a true crisis of insolvency. The problem for policy makers is to diagnose which is taking place. If the bank is fundamentally sound, lending it some cash or other liquid assets will allow the panic to subside and confidence to be restored. On the other hand, if the bank is truly insolvent, temporary loans will not help it. Its assets are less than its liabilities and more drastic action is required – donate enough government funds to make it solvent again, force shareholders to do the same or close the bank.

In a **liquidity crisis**, an institution is temporarily unable to meet immediate requests for payment even though its underlying assets exceed its liabilities.

When the financial crisis first erupted in the UK in 2008, with a panic by depositors of Northern Rock and long queues of people shown on the news, this was initially interpreted as a liquidity crisis that some temporary assistance, or government promises to guarantee depositors' money, could reverse. It soon became apparent that the Rock was in much deeper trouble than that. Its entire solvency was at stake.

Whether it is a liquidity or a **solvency crisis** determines which arm of government might potentially be involved in a solution. The Bank of England can make emergency loans if it expects them to be repaid in full, for there are then no long-term issues for taxpayers. Fixing insolvency, on the other hand, requires permanent injections of taxpayers' money, for which the authority of the Treasury (and ultimately the prime minister) is required. Since crises move quickly once they have begun, co-ordination of the fiscal and monetary authorities is vital if crisis resolution is to be effective.

In a **solvency crisis**, an institution's assets have become less than its liabilities. It is bankrupt without a rapid new injection of assets from government or shareholders.

CASE 18.1

THE SUB-PRIME CRISIS AND ITS AFTERMATH

Most countries experienced an explosion in house prices circa 2005-06. Inflation appeared to have been conquered, interest rates were low and borrowing did not look too risky. Many of those working in the financial sector – whether in banks or in property – were getting big bonuses based on the volume of lending not the prudence of these loans. There were strong incentives to dream up new products and find new lines of business.

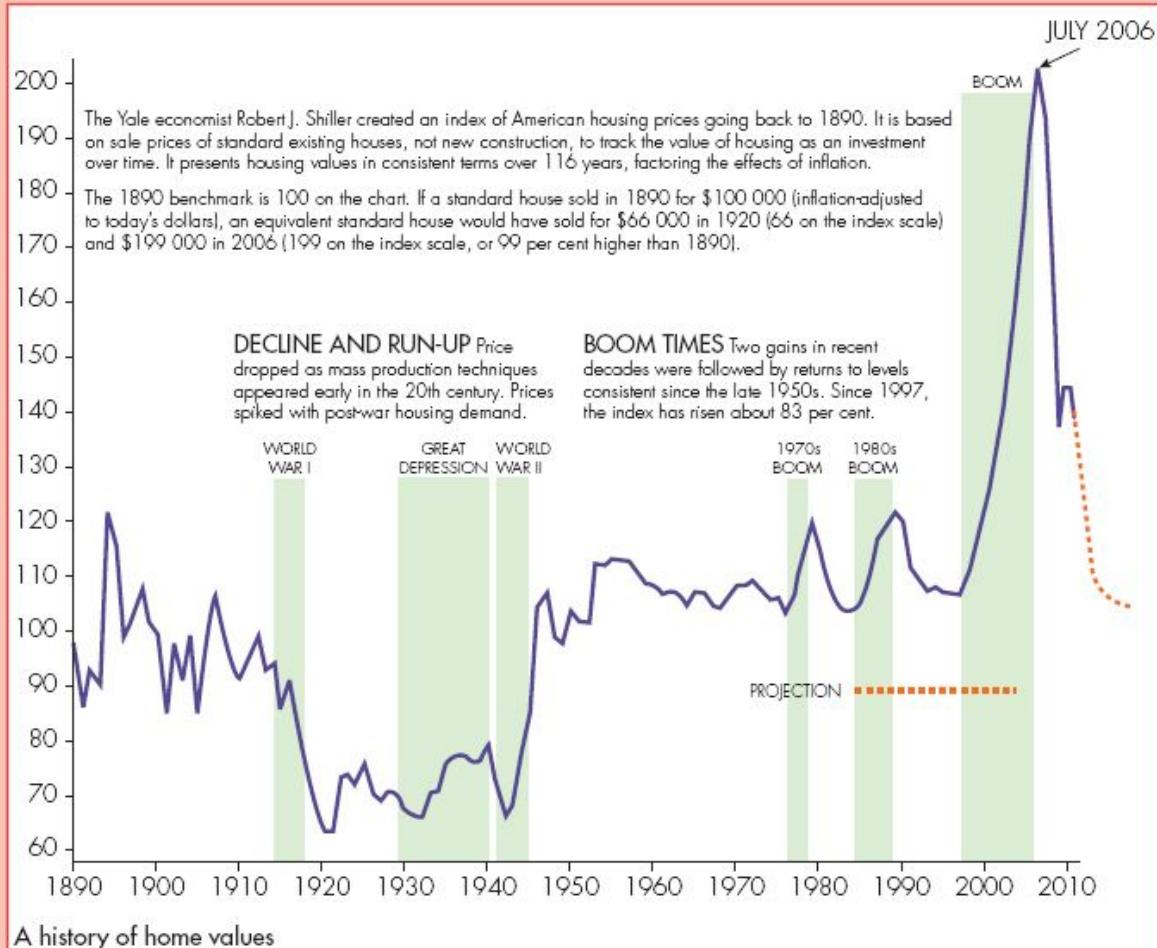
In the US, one of these new products was the sub-prime mortgage, a housing loan to a low-income high-risk person who had previously been unable to borrow in order to buy a house. Most of these mortgages were at variable interest rates: although initially low and 'affordable', they were subsequently raised if either general market interest rates rose or if lots of people started to default and it became necessary to build a larger risk premium into the interest rate. It is unclear how much of this was explained to the low-income people being signed up for first-time mortgages.

US house prices peaked in 2006. As they then fell, lenders got scared and began to raise mortgage interest rates, driving many of the poor to default. Suddenly, these sub-prime mortgages were worth a lot less than had been thought. And the crisis fed upon itself. The more scared people became, the more asset prices fell, validating the initial fears.

If mortgages had simply been issued by a few institutions specializing in loans for house purchase, the damage might have been quarantined. The US government would have had to decide whether to (a) let these particular institutions go bust or (b) inject taxpayers' money to prop them up.

Securitization transformed a local crisis into a global problem. Smart financiers, driven by the prospect of new business and big personal bonuses, had grouped lots of individual sub-prime mortgages into large bundles and sold them on to new buyers in London, Frankfurt and Mumbai. The market was convinced that this trick was a bit like insurance – one poor sub-prime household might go bust, but they would not all go bust together. Holding a large bundle made them safer, just as an insurance company pools the risk of individual burglary by having a large number of clients. This was the alchemy of risk reduction, a recipe for immediate profits and bonuses. Two things went wrong. First, buyers of securitized mortgages had miscalculated. Unsurprisingly, all sub-prime borrowers got into trouble at the same time, as a result of a fall in house prices, a fall in confidence and a rise in risk perception. Supposedly smart bankers in London, New York and other financial capitals had mispriced the risk: the securitized bundles were riskier than had been thought.

The figure shows US house prices, inflation-adjusted, since 1890. The extent of the house price bubble was unprecedented by 2006, as was the severity of the subsequent crash – it was not until 2012 that US house prices began to turn upwards again.



Sources: Robert J. Shiller, *Irrational Exuberance*, 2nd Edition, 2006; Bill Marsh/The New York Times; Steve Barry for The Big Picture, 2011.

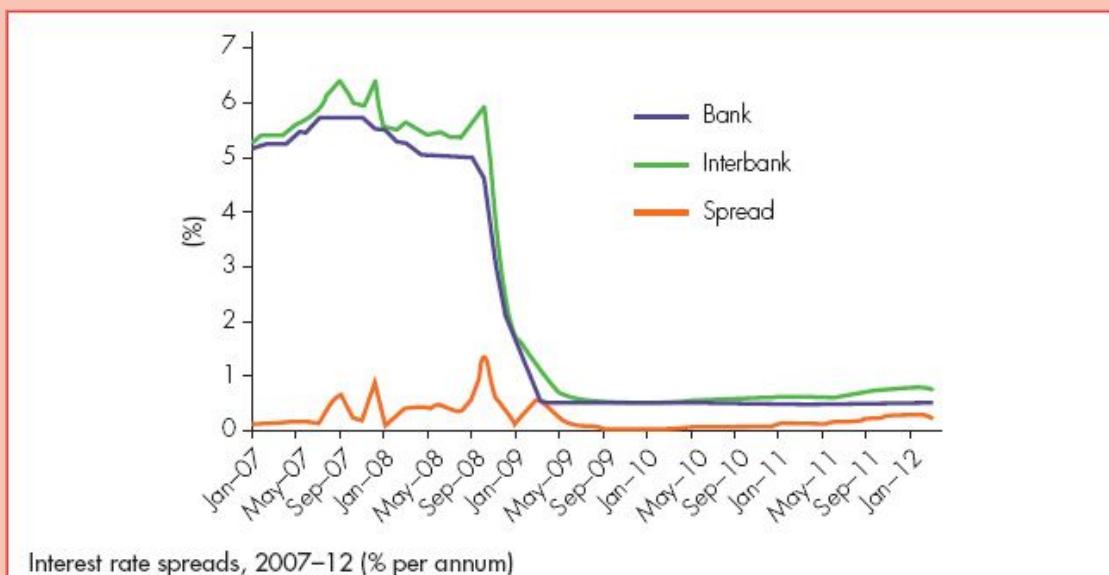


Sources: Nationwide Building Society; <http://www.landregistry.gov.uk/houseprices>.

The figure above shows real house prices in the UK, which collapsed a couple of years later, in 2008. Again there was a severe correction, though less acute than in the US.

Second, the perfect storm did indeed arise. As US house prices fell sharply, the chain of events was triggered. Banks found their assets worth much less than they had thought. Worse, the boards of the banks had not even realized the extent to which their bonus-hungry employees had exposed them to such large risks.

As the solvency of banks came into question, people became reluctant to lend to banks, and banks themselves became reluctant to lend to anyone else – aware of the potentially fatal hole in their balance sheet, banks prioritized using resources to rebuild their own reserves. The entire, apparently well-oiled, system of liquidity dried up as banks disappeared from the lending business. One way to see how dramatic this was is to examine interest rate spreads; that is, the difference between the interest rate banks were charging for the few scarce loans they were prepared to make and the interest rate at which banks could borrow from the Bank of England.



Source: www.bankofengland.co.uk.

The interest rate spreads figure plots in purple the official bank rate, the interest rate at which commercial banks can borrow from the Bank of England; and in green plots the interbank rate at which banks lend temporary excess funds to one another. In normal times, competition between banks means that the ‘profit margin’ between the lending rate and the borrowing rate is very small. Banks charging too much are quickly outcompeted by banks prepared to offer a better deal. This spread between the interbank rate at which banks lend and the bank rate at which they can borrow from the Bank of England is shown in orange.

The interest rate spread is a thermometer with which to monitor the health of the banking system. In a healthy system, banks feel confident, competition prevails and spreads are very small. When a crisis breaks out, spreads shoot up, raising the price to ration loans only to the very safest customers. And the volume of bank lending collapses. The figure above shows spreads rising in 2007 and more sharply in 2008-09. Notice that by September 2009 the crisis was largely over,

though new concerns (based on possible defaults within the Eurozone) re-emerged in 2011-12.

The chronology of crisis (through UK spectacles)

2006	US house prices start to fall, the sub-prime crisis begins, interest spreads edge up around the world, bank lending slows down, liquidity begins to evaporate.
2007	UK bank Northern Rock hits a liquidity crisis in September 2007 – not yet because its asset values have fallen (UK house prices are still rising) – but because UK credit markets have dried up and the Rock cannot roll over its short-term loans. Borrowing short term to lend long term for housing loans is always a risky business. Crisis temporarily resolved once Bank of England agree to provide liquidity financing to Northern Rock.
2008	February – with UK house prices now having peaked, the market becomes worried not just about Northern Rock's ability to refinance its loans, but also about the value of its underlying assets. A full-blown insolvency crisis. UK government decides to nationalize the Rock. March – US investment bank Bear Sterns, a pioneer in securitizing mortgage-backed securities, suffers an insolvency crisis. Competitor JPMorgan Chase buys Bear Sterns cheaply. Getting competitors to take over failing banks is often a good way out, since no bankruptcy or severe dislocation ensues. September – US Treasury has to bail out Freddie Mac and Fannie Mae, the two largest mortgage lenders in the US. – US investment bank giant Lehman Brothers is allowed to go bankrupt without US Treasury managing to arrange a satisfactory bail out, arguably the single event that triggered financial panic around the world, from which no country was immune. October – Royal Bank of Scotland, which had overextended itself buying Dutch bank ABN AMRO at too high a price after a bidding war with Barclays, faces a solvency crisis, temporarily resolved by UK Treasury taking a 58 per cent stake in RBS.
2009	January – UK government persuades Lloyds bank to buy the potentially insolvent Halifax Bank of Scotland group. Lloyds' shareholders subsequently discover HBOS worth much less than they paid for it. The outcome is good for the system but bad for Lloyds. UK taxpayer raises stake in RBS to 84 per cent to prevent its bankruptcy. Governments around the world gradually admit to the scale of government injections to bail out their banks. Since taxes are not raised to pay for this, the initial consequence is a huge jump in levels of government deficits and government debt.
	Bank behaviour is becoming more cautious – they are reluctant to lend to anyone who is not guaranteed to repay, and they are charging customers high interest rates to make profits to rebuild their own balance sheets. Governments are discussing changes in the regulation of banks to ensure greater control and less risk taking.
2011	Having split Northern Rock into a 'good bank' and a 'bad bank', UK government agrees sale of good bank, and its associated high street branches, to Virgin Money. Still unable to sell RBS. A commission chaired by Sir John Vickers, Oxford economics professor and former chief economist of the Bank of England, recommends tackling the problem of banks being 'too big to fail' and hence holding the government hostage in a crisis. The report stopped short of recommending that 'boring retail banking' and 'casino investment banking' be undertaken in separate companies, entailing the splitting-up of existing banks, but did propose that retail banking and investment banking be separately ring-fenced within the company, and any implicit government guarantee being extended only to the retail banking part intrinsic to protection of the payments system. Private creditors of Greece are strong-armed into 'volunteering' to take a substantial writedown of the value of their loans. Mario Draghi, new governor of the European Central Bank, provides massive medium-term loans to European banks, preventing a lack of liquidity leading to another major crisis. This buys time and is hailed a success. However, European governments continue to dither in providing a fundamental solution to the solvency problems of banks/governments, the two having become inseparable.
2012	The Eurozone crisis rumbles on, all European banks discovering they have more exposure than they thought to a default by governments of peripheral Eurozone countries such as Greece. Even banks with little direct lending to the Greek or Spanish government find they have lent to French or other country banks that in turn have lent extensively to peripheral countries. In July, Mario Draghi promised to do 'whatever it takes' to solve the crisis, and European stock markets boom. UK government agrees in principle to adopt the reforms in the Vickers report and implement them 'by 2019'. Critics argue implementation should be rapid in case there is another

crisis; bankers argue that London will be disadvantaged if sweeping restrictions are quickly imposed.

Summary

- Money has four functions: a **medium of exchange** or means of payment, a **store of value**, a **unit of account** and a **standard of deferred payment**. Its use as a medium of exchange distinguishes money from other assets.
- In a **barter economy**, trading is costly because there must be a double coincidence of wants. Using a medium of exchange reduces the cost of matching buyers and sellers, letting society devote scarce resources to other things. A **token money** has a higher value as a medium of exchange than in any other use. Because its monetary value greatly exceeds its production cost, token money economizes a lot on the resources needed for transacting.
- Token money is accepted either because people believe it can subsequently be used to make payments or because the government makes it legal tender. The government controls the supply of token money.
- Banks **create money** by making loans and creating deposits that are not fully backed by cash reserves. These deposits add to the medium of exchange. Deciding how many reserves to hold involves a trade-off between interest earnings and the danger of insolvency.
- Modern banks attract deposits by acting as **financial intermediaries**. A national system of clearing cheques, a convenient form of payment, attracts funds into sight deposits. Interest-bearing time deposits attract further funds. In turn, banks lend out money as short-term liquid loans, as longer-term less liquid advances or by purchasing securities.
- Sophisticated financial markets for short-term liquid lending allow modern banks to operate with very low cash reserves relative to deposits. The **money supply** is currency in circulation plus deposits. Most is the latter.
- The **monetary base M0** is currency in circulation plus banks' cash reserves. The **money multiplier**, the ratio of the money supply to the monetary base, is big. The money multiplier is larger (a) the smaller is the desired cash ratio of the banks and (b) the smaller is the private sector's desired ratio of cash in circulation to deposits.

- Financial deregulation has allowed building societies into the banking business. **M4** is a broad measure of money and includes deposits at both banks and building societies.
- The **demand for money** is a demand for real money, for its subsequent purchasing power over goods. The demand for **narrow money** balances the transactions and precautionary benefits of holding another pound with the interest sacrificed by not holding interest-bearing assets instead. The quantity of real money demanded falls as the interest rate rises. Higher real income raises real money demand at each interest rate.
- For **wide money** such as M4, the asset motive for holding money also matters. When other interest-bearing assets are risky, people diversify by holding some safe money. With no immediate need to transact, this leads to an asset demand for holding interest-bearing bank deposits. This demand is larger, the larger the total wealth to be invested and the lower the interest differential between deposits and risky assets.

Review questions



EASY

- 1 (a) A person trades in a car when buying another. Is the used car a medium of exchange? Is this a barter transaction? (b) Could you tell by watching someone buying mints (white discs) with coins (bronze discs) which one is money?
- 2 **Common fallacies** Why are these statements wrong? (a) Since their liabilities equal their assets, banks cannot create anything. (b) The money supply has risen because of tax evasion. Since cash is untraceable, people are putting less in the banks.
- 3 Saying that banks have become too big to fail means: (a) large banks are safer, (b) large banks are less safe, (c) managers of large banks realize they can take risks because politicians will have to bail them out if things go wrong?
- 4 Suppose sight deposits = £30, time deposits = £60, banks' cash reserves = £2, currency in circulation = £12, building society deposits = £20. Calculate M0 and M4.
- 5 Which of these is the correct answer? After the financial crash, bank lending to the private sector slumped because (a) new regulations were introduced, (b) banks were broken up to prevent another crisis, (c) banks thought prospective borrowers were too risky, (d) the value of bank reserves had fallen, (e) only answers a and d, (f) only answers c and d.

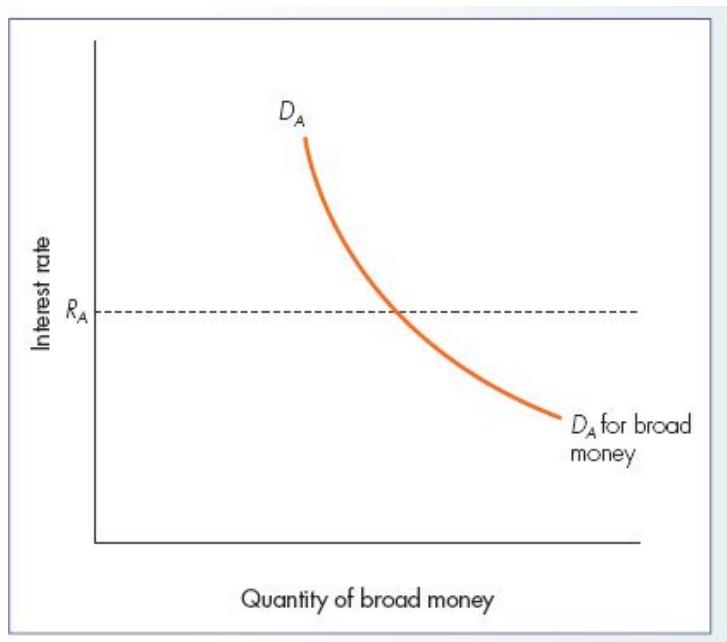
MEDIUM

- 6 Initially gold coins were used as money but people could melt them down and use the gold for industrial purposes. (a) What must have been the relative value of gold in these two uses? (b) Explain the circumstances in which gold could become a token money. (c) Explain the circumstances in which gold could disappear from monetary circulation completely.
- 7 Suppose banks raise interest rates on time deposits whenever interest rates on bank loans and other assets rise. Does a rise in the general level of interest rates have a big or small effect on the demand for time deposits?
- 8 How do commercial banks create money?
- 9 Suppose banks initially wish to hold reserves R equal to 1 per cent of the deposits D that they provide, and that the general public wishes to hold cash C equal to 2 per cent of the deposits that they hold. The monetary base $H = C + R$, and broad money is cash with the public plus bank reserves. What is the value of the money multiplier?

MEDIUM

- 10 Since credit cards can be used to make payments, why are they not treated as money?
- 11 Would it make sense to include (a) travellers' cheques, (b) student rail cards or (c) credit cards in measures of the money supply?
- 12 Essay question Lots of institutions accept deposits and reissue them on demand – building societies, Christmas savings clubs and theatre cloakrooms. What is the key feature of banks that distinguishes them from other institutions? Why does this matter?
- 13 The following diagram shows the downward-sloping demand curve for broad money. (a) Suppose the supply curve is vertical. Show the money market equilibrium. (b) Depict the effect of the financial crash of 2008 on supply of and demand for broad money. (c) Why did monetary policy create substantial quantities of narrow money in these circumstances?

HARD



14 In the diagram in Question 13, what would be the consequence of a sharp increase in confidence about the health of the financial sector? How would monetary policy be likely to respond?

- 1 For an interesting account of cigarettes as money in prisoner-of-war camps, see R. A. Radford, ‘The economic organisation of a POW camp, *Economica* 48 (1945): 189–201, which was introduced in Activity 1.1.
- 2 The existence of forgers confirms society is economizing on scarce resources by producing money whose value as a medium of exchange exceeds its production cost.
- 3 Society continues to find new ways to save scarce resources in producing and using a medium of exchange. Many people use credit cards. Some supermarket tills directly debit customers’ bank accounts. And shopping via the TV, telephone and Internet is growing rapidly.
- 4 In fact, building societies now issue cheque books to their depositors, which is why building societies are now included in monetary statistics.
- 5 By allowing us to pay all at once when the statement arrives monthly, credit cards have this effect.
- 6 The cost of holding money is the differential return between bonds and money. If p is the inflation rate and r the nominal interest rate, the real interest rate is $r - p$. In financial terms, the real return on money is $-p$, the rate at which the purchasing power of money is eroded by inflation. The differential real return between bonds and money is $(r - p) - (-p) = r$. The *nominal* interest rate is the opportunity cost of holding money.

CHAPTER 19

Interest rates and monetary transmission

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 realise how a central bank can affect the money supply
- 2 understand quantitative easing
- 3 describe the central bank's role in financial regulation
- 4 describe money market equilibrium
- 5 recognise an intermediate target for monetary policy
- 6 understand the transmission mechanism of monetary policy
- 7 describe how a central bank sets interest rates
- 8 understand how interest rates affect consumption and investment demand

Today, every country of any size has a central bank. Originally private firms in business for profit, central banks came under public control as governments placed more emphasis on monetary policy. Founded in 1694, the Bank of England (www.bankofengland.co.uk) was not nationalized until 1947. The Federal Reserve System, the US central bank, was not set up until 1913. Within the Eurozone, individual central banks survive, but the European Central Bank, established by the 1998 Treaty of Amsterdam, is in charge of the single monetary policy.

A **central bank** is banker to the government and to the banks. It conducts monetary policy.

This chapter examines the role of the **central bank**, and shows how it influences financial markets. The central bank influences the supply of money. Combining this with the demand for money, examined in the previous chapter, we analyse money market equilibrium. The central bank's monopoly on the supply of cash allows it to control equilibrium interest rates. Finally, we discuss how monetary policy decides what interest rates to set.

19.1 The Bank of England

The Bank of England, usually known simply as the Bank, is the UK central bank. It is divided into Issue and Banking Departments. Its balance sheet is shown in Table 19.1.

Banknotes and coins are liabilities of the Bank. Your £1 coin is a debt of the Bank of England. When commercial banks have reserves at the Bank of England, these are owed by the Bank to the commercial banks, the ultimate owners of these reserves. Cash in circulation, banks' reserves held at the Bank of England and other liabilities of the Bank in total came to £414 billion in November 2012.

What were the corresponding assets? To introduce cash into circulation, the Issue Department engages in open market operations to buy financial securities issued by the government, commercial firms or local authorities. One such asset is bonds; another is a reverse repo, which we explain below. In 2012 the largest class of assets of the Bank was in fact 'other assets', which were largely explained by the programme of 'quantitative easing' described in Concept 19.1. Together, these assets came to £414 billion.

Table 19.1 Bank of England, balance sheet, November 2012

Liabilities	£bn	Assets	£bn
Cash in circulation	57	Reverse repos	11
Banks' reserves	282	Bonds bought	14
Other liabilities	75	Other assets	389
Total liabilities	414	Total assets	414

Source: Data taken from Statistical Interactive Database, © Bank of England, <http://www.bankofengland.co.uk/boeapps/iadb> accessed on 18/06/2013.

Table 19.1 resembles the balance sheet of a commercial bank, with one key difference. *A central bank cannot go bankrupt.* You take £50 to the

Bank and cash it in for £50. The Bank gives you £50 in cash. It can always create new cash. Hence, it can never run out of money.

A **repo** is the sale of an asset with a simultaneous agreement to repurchase later.

A **reverse repo** is a purchase with a simultaneous agreement to resell later.

A **repo** is a *sale and repurchase agreement*. A bank sells you a bond, simultaneously agreeing to buy it back at a specified price on a particular future date. You have made the bank a short-term loan secured or ‘backed’ by the long-term bond temporarily in your ownership. Thus repos use the outstanding stock of *long-term* assets as backing for new and secured *short-term* loans.

One party’s repo is the other party’s **reverse repo**. Suppose you get a short-term loan from the bank by initially selling bonds to the bank, plus an agreement for you to repurchase the bonds at a specified date in the near future at a price agreed now. Reverse repos are effectively secured temporary fixed-term loans by the Bank. That is why they appear on the asset side of its balance sheet.

Repos and reverse repos are very like other short-term lending and borrowing. The Bank of England used to alter cash in circulation by buying or selling Treasury bills. Now it follows other central banks in using the repo market to conduct these ‘open market operations’ in order to alter cash in circulation.

Similarly, Table 19.2 shows the balance sheet of the *Eurosystem*, which comprises the European Central Bank and the 17 central banks of Eurozone member states.

Table 19.2 Eurosystem balance sheet, March 2012

Liabilities	€bn	Assets	€bn
Cash in circulation	871	Gold	423
Banks' reserves	1149	Loan to Eurobanks	1130
Other liabilities	75	Financial securities	632
		Other assets	838
Total liabilities	3023	Total assets	3023

Source: Data taken from European Central Bank, ‘Consolidated financial statement of the Eurosystem as at 2 March 2012’, www.ecb.int/press/pr/wfs/2012/html/fs120306.en.html.

19.2

Traditional means of monetary control

The **money supply** is currency in circulation *outside* the banking system, plus deposits of commercial banks and building societies.

The **money supply**(M4) is partly a liability of the Bank (currency in private circulation) and partly a liability of banks (bank deposits). Henceforth, we talk of ‘banks’ without distinguishing between banks and building societies.

The central bank can therefore affect broad money M4 either by affecting the cash in circulation or by affecting the number of deposits for any given amount of cash in circulation. We begin with policies that affect the latter.

Reserve requirements

A **required reserve ratio** is a minimum ratio of cash reserves to deposits that banks are required to hold.

Banks can hold more than the required cash reserves but not less. If their reserves fall below the required amount, they must immediately borrow cash, usually from the central bank, to restore their **required reserve ratio**.

Suppose banks have £1 billion in cash and, for commercial purposes, want cash reserves equal to 5 per cent of deposits. Deposits are 20 times cash reserves. Banks create £20 billion of deposits against their £1 billion cash reserves. However, if there is a reserve requirement of 10 per cent, banks only create £10 billion deposits against cash reserves of £1 billion. The money supply falls from £20 billion to £10 billion.

When the central bank imposes a higher reserve requirement than the reserve ratio that prudent banks would anyway have maintained, the effect is fewer bank deposits and a lower money supply for any amount

of cash in circulation. Raising the reserve requirement reduces the money supply.

The discount rate

The **discount rate** is the interest rate that the Bank charges when banks want to borrow cash.

Suppose banks think the *minimum* safe ratio of cash to deposits is 10 per cent. It does not matter whether this figure is a commercial judgement or a requirement imposed by the Bank. Banks may also hold extra cash. If their cash reserves are 12 per cent of deposits, how far dare they let their cash fall towards the 10 per cent minimum?

Banks balance the interest rate on extra lending against the cost incurred if withdrawals push their cash reserves below the critical 10 per cent. If the central bank lends to banks at market interest rates, there is no penalty incurred from being caught short and having to borrow from the central bank. Banks lend as much as they can and their cash reserves fall to the minimum required.

Suppose the Bank only lends to banks at an interest rate above market interest rates. Now commercial banks will not drive down their reserves to the minimum permitted. They hold extra cash as a cushion, to avoid possibly having to borrow from the central bank at penalty rates.

By setting the discount rate above general interest rates, the Bank can induce banks voluntarily to hold extra cash reserves. Bank deposits are a lower multiple of banks' cash reserves, and the money supply is lower for any given level of cash in circulation. Variations in the discount rate can change the money supply.

Open market operations

Whereas the previous two methods of monetary control alter the amount of deposits created for any given amount of cash in circulation, open market operations alter the amount of cash in circulation. Since this then affects the amount of deposits that banks wish to create, open market operations alter the money supply both directly (via the effect on cash in

circulation) and indirectly (via the induced effect on the number of deposits created).

An **open market operation** occurs when the central bank alters the monetary base by buying or selling financial securities in the open market.

The Bank prints £1 million of new banknotes and buys bonds on the **open market**. There are £1 million fewer bonds in private hands but £1 million more in cash. Some of the extra cash is held in private circulation but most is deposited with the banks, which then expand deposit lending against their higher cash reserves. Conversely, if the Bank sells £1 million of bonds from its existing holdings, the monetary base falls by £1 million. Banks lose cash reserves, have to reduce deposit lending and the money supply falls.

Open market operations are nowadays the principal channel by which the central bank affects the money supply. Having discussed the central bank's role in monetary control, we turn next to its role in financial stability.

19.3 Lender of last resort

Modern fractional reserve banking lets society produce the medium of exchange with tiny inputs of scarce physical resources. But the efficient production of the medium of exchange yields a system of fractional reserve banking vulnerable to financial panics. Since banks have too few reserves to meet a withdrawal of all their deposits, a hint of big withdrawals may become a self-fulfilling prophecy as people scramble to get their money out before the banks go bust.

In Chapter 18 we described how the central bank can create and lend cash to banks to stave off a liquidity crisis. This requires a guarantee that banks can get cash if they really need it. The central bank is the only institution that can manufacture cash in unlimited amounts. The threat of financial panics is greatly diminished if it is known that the central bank will act as **lender of last resort**. As lender of last resort, the Bank can maintain confidence in the banking system, provided the underlying solvency of banks is not threatened. What went wrong during the financial crisis was that some banks acquired assets that turned out to be

worthless, and became insolvent. Last resort lending could not save them since it made no difference to their underlying solvency.

The **lender of last resort** lends to banks when financial panic threatens the financial system.

Prudential regulation

The prospect of insolvency raises two issues: how to respond to a particular insolvency crisis, and how to prevent such a crisis arising in the first place.

A **capital adequacy ratio** is a required minimum value of bank capital relative to its outstanding loans and investments.

Generally, it is the shareholders of the particular bank that bear the cost of its poor performance. To try to make sure that shareholders have sufficient funds for this purpose, financial regulations require banks to meet **capital adequacy ratios**.¹

Banks face the liquidity risk that depositors may withdraw money before banks can sell their less liquid assets, and the solvency risk that a downward revaluation of the value of their assets may leave assets worth less than their liabilities. Bank reserves help protect against liquidity risk. Bank capital helps protect against solvency risk. Bank capital is supplied originally by shareholders or represents bank profits ploughed back into the business.

A crisis depletes this capital reserve and thereby reduces the share price of the bank. *Shareholders* suffer, but *depositors* are protected if the bank still has adequate bank reserves to meet the prospect of future withdrawals. Depositors may also have an explicit or implicit guarantee from the government.

If a bank makes larger losses, it may go bankrupt. Losses incurred by rogue trader Nick Leeson brought down Barings Bank in the 1990s. Typically, governments then compensate depositors but not shareholders. Barings was actually sold to Dutch bank ING for a notional amount and deposits were honoured in full. The knowledge that depositors are unlikely to suffer helps prevent unjustified financial panics. The

knowledge that shareholders *are* likely to suffer helps keep management on its toes.

Three things went wrong in the perfect storm of 2008/09. First, the magnitude of the initial shock was very large. Greedy banks had borrowed billions to speculate on securitized products whose true risk characteristics they did not properly understand, and which subsequently proved a very bad investment.

Second, capital adequacy regulations had been poorly designed. What was adequate financial backing by shareholders in good times turned out to be grossly inadequate capital reserves in a big crisis. This has led to calls for future capital adequacy requirements to be variable – less onerous when economies are doing well, but increasingly demanding as risks of crises get larger. As the sub-prime crisis got off the ground in 2006, variable capital adequacy requirements would have required banks around the world to retain more profits or ask shareholders for new funds in order to build up capital reserves.

The third lesson is that many banks had become ‘too big to fail’. In a capitalist economy, you might have expected insolvent banks to be made bankrupt in the same way as a defunct car company or steel producer. However, as the US discovered when it allowed Lehman Brothers to go bust, if the bank is large enough it causes massive ripples throughout the financial system. Sometimes, the government concludes that injecting taxpayers’ money into keeping the bank going is the lesser of two evils, and preferable to letting the bank go under.

This, of course, is what happened in many Western economies in 2008. Nor were governments prepared instantly to raise taxes to pay for this huge spike in their spending on bank bailouts. Instead, they borrowed money and acquired debt; as a result they now owe huge amounts that will take years to pay back.

How do we minimize the chances of such an awful dilemma arising in the future? Case 19.1 discusses possible solutions.

CASE 19.1

PREVENTING A FUTURE BANKING CRISIS?

In our discussion of microeconomics in Part Two, we examined two important ideas: moral hazard and imperfect commitment. *Moral hazard* arises when the adoption of a set of rules that would be ideal under perfect information then fosters unwelcome behaviour because it is too costly subsequently to monitor individual behaviour. For example, insurance reduces the cost of bearing risk, which is a good thing. However, fully insured people may no longer bother to act prudently to avoid risk, which is the consequence of moral hazard. It is too costly to verify which individuals had acted prudently and which had not. We generally solve this problem with a compromise: we offer partial but not full insurance so that individuals still have an incentive to act prudently.

Bank of England governor, Mervyn King, has repeatedly drawn attention to the moral hazard problems in bailing out banks. Even if bank bailouts are helpful in preventing a crisis from escalating today, the signal that future bailouts are also likely may increase behaviour that makes future crises more likely. If all bankers know they are going to be bailed out, they might as well take big risks. When these come off, the banks do well and bankers' bonuses are high; when the risks prove disastrous, the government steps in. This is a one-way bet for bankers.

Imperfect commitment is also a problem. Most governments say they will be tough in the future – thereby scaring bankers into more prudent behaviour today – but if bankers can deduce that governments always cave in when it comes to the crunch, tough words today are empty threats that the bankers can ignore.

Either we live with the reality that bailouts are likely, because the financial tsunami caused by allowing Lehman Brothers to go bankrupt can never be repeated, or we have to find a structural solution that prevents such problems recurring in the future.

The first possibility is the separation of *retail banking* – the relatively boring business of taking deposits from the general public and making loans of a traditional nature, a key aspect of which is that these banks are banned from speculating with their own capital – and *investment banking*, in which banks may gamble extensively with their own funds in derivatives and other volatile and fancy products. This distinction is sometimes called that between *narrow* banks and *broad* banks.

We have seen all this before. In response to the Wall Street crash of 1929, after a previous episode of banking irresponsibility, the US passed the Glass–Steagall Act, which prevented retail banks undertaking risky investment banking activities. The intention was to ensure that neither the real economy nor government finances would again be endangered by casino banking. If investment banks got into trouble, they could be allowed to fail and repercussions for the rest of us would be much smaller. In turn, this made it more credible that they would indeed be allowed to fail if necessary.

For 70 years this approach was largely a success. Yet as the financial sector became increasingly competitive internationally, institutions facing legal restrictions on their behaviour pressed to be allowed to join the lucrative investment banking business which had prospered during decades of stability and growth. The UK began liberalizing its financial sector in 1986, the so-called Big Bang that paved the way for building societies to behave like banks, and for banks to behave like investment banks. Gradually, other European countries were also forced to liberalize in order to get a share of the lucrative investment banking business. The US repealed the Glass–Steagall Act in 1999.

The first solution affects the permitted structure of banks; a second possible solution limits the types of activity that deposit-taking retail banks can undertake. In 2010 the US adopted the Volcker Rule, prohibiting retail deposit banks from speculating with their own capital. In the UK, the Vickers Commission recommended in 2011 not a complete separation into retail and investment banks but a clear separation of these activities within each bank. Retail activities could still rely on government support in a crisis. Investment activities would be the responsibility of the bank's shareholders and allowed to fail if necessary. The UK has committed to implement this proposal by 2019, leaving a long intervening period in which its banks will still be considered vulnerable.

A third solution is to let all banks undertake all types of transaction, but to place an absolute limit on the size of banks eligible for deposit guarantees and fiscal bailouts. If the problem is that some banks became 'too big to fail' – governments then being forced to rescue them – the solution is to keep banks sufficiently small that it

is credible that their failure would not trigger automatic bailout by government.

It is impossible for regulators to prevent business failure, and undesirable to pursue that objective. The essential dynamic of the market economy is that good businesses succeed and bad ones do not. There is a sense in which the bankruptcy of Lehman was a triumph of capitalism, not a failure. It was badly run, it employed greedy and overpaid individuals, and the services it provided were of marginal social value at best. It took risks that did not come off and went bust. That is how the market economy works.

The problem now is how to have greater stability while extricating ourselves from the ‘too big to fail’ commitment, and taking a realistic view of the limits of regulation. ‘Too big to fail’ exposes taxpayers to unlimited, uncontrolled liabilities. The moral hazard problem is not just that risk-taking within institutions that are too big to fail is encouraged but that private risk-monitoring of those institutions is discouraged.

John Kay, ‘Too big to fail is too dumb an idea to keep’, *Financial Times*, 27 October 2009.

A fourth possibility is to rely on stronger prudential supervision by regulatory agencies, particularly by enforcing tougher capital adequacy ratios. If banks have larger reserves, failure of private banks is less likely. Banks are forced to set aside large reserves that earn little return but are available as shareholder capital with which to meet future crises.

One problem is the need to co-ordinate regulation across different regulators. When RBS took excessive risks, should this have been a concern for the Financial Services Authority (the UK body charged with supervising financial institutions), the Bank of England (the potential lender of last resort in a liquidity crisis) or the Treasury (the government department potentially responsible for injecting taxpayers’ money in a solvency crisis)?

This problem is particularly acute within the Eurozone. When a bank operates across many countries, whose taxpayers pay if the bank fails? This dilemma has led to demands for a *banking union* within the Eurozone, common regulations and clear criteria about

what happens in the event of failure. As is so often the case, the economics is easier than the politics. Politicians find it hard to convince national voters to ‘give up’ national sovereignty, even if pooling sovereignty might lead to a much more effective outcome.

Thus, it may be easier to obtain a one-off agreement for a long-term structural solution, as in Glass–Seagall, than to co-ordinate different agencies on a daily basis to evaluate ongoing performance.

Moreover, much of this risky financial business is globally footloose. If some financial centres regulate more than others, private business may tend to migrate to the least-intrusive location. Competition *between* financial centres was and remains part of the problem. London might have regulated earlier if it had been less frightened of losing business to Frankfurt and New York. This suggests that any reforms that will make an enduring difference may have to be negotiated at the level of the top ten global countries, not merely a country at a time.

Basel III Accord

The third Basel Accord is a global, voluntary, regulatory standard on bank capital adequacy, stress testing and market liquidity risk. During 2014–, tougher capital requirements for banks will gradually be phased in. During 2013–17, measures of risk will be tracked for individual banks, and passing these tests will become mandatory by 2018. During 2015–18, minimum liquidity requirements will steadily be phased in. All major countries are expected to endorse Basel III. Its provisions mark a considerable step away from the idea that banks need little regulation.

19.4

Equilibrium in financial markets

Having discussed the role of the central bank in financial crises, we now revert to its more normal role.

The traditional account of central banking views the central bank as controlling the *nominal* money supply (it has monopoly power to supply narrow money; that is, cash plus banks’ reserves at the central bank). If, but only if, the money multiplier is stable, this allows it to control the nominal supply of broad money by setting the quantity of narrow money.

When we simplify by assuming that the price of goods is fixed, the central bank also controls the *real* money supply. In later chapters, we allow the price level to change. Changes in nominal money tend to lead to changes in prices. The central bank can still control the **real money supply** M/P in the short run – it can change M faster than prices P respond – but, in the long run, other forces determine real money M/P . For the moment, we treat the price level as fixed.

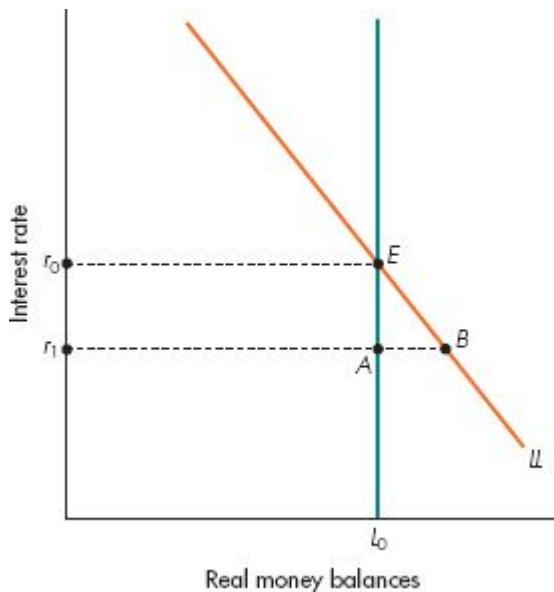
The **real money supply** L is the nominal money supply M divided by the price level P .

In the previous chapter, we argued that the quantity of real money demanded rises when real income rises, but falls when the nominal interest rate rises.

Money market equilibrium

In **money market equilibrium** the quantity of real balances demanded and supplied is equal.

Figure 19.1 shows the demand curve LL for real money balances for a given real income. The higher the interest rate and the cost of holding money, the less real money is demanded. With a given price level, the central bank controls the quantity of nominal money and real money. The supply curve is vertical at this quantity of real money L_0 . Equilibrium is at E . At the interest rate r_0 , the real money people wish to hold just equals the outstanding stock L_0 .



The demand schedule LL is drawn for a given level of real income. The higher the opportunity cost of holding money, the lower the real balances demanded. The real money supply schedule is vertical at L_0 . The equilibrium point is E and the equilibrium interest rate r_0 . At a lower interest rate r_1 there is excess demand for money AB . There must be a corresponding excess supply of bonds. This reduces bond prices and increases the return on bonds, driving the interest rate up to its equilibrium level at which both markets clear.

Figure 19.1 Money market equilibrium

Suppose the interest rate is r_1 , below the equilibrium level r_0 . There is excess demand for money AB in Figure 19.1. How does this excess demand for money bid the interest rate up from r_1 to r_0 to restore equilibrium? The answer is rather subtle. Strictly speaking, there is no market for money. Money is the medium of exchange for payments and receipts in *other* markets. A market for money would exchange pounds for pounds.

The other market relevant to Figure 19.1 is the market for bonds. Since the interest rate is the cost of holding money, people who do not hold money hold bonds. What happens explicitly in the market for bonds determines what is happening in the implicit market for money in Figure 19.1.

Real wealth W is the existing supply of real money L_0 and real bonds B_0 . People divide their wealth W between desired real bond holdings BD and desired real money holdings LD . Hence

$$B_0 - B_D = L_D - L_0$$

An excess demand for money must be exactly matched by an excess supply of bonds. Otherwise people are planning to hold more wealth than they actually possess.

In Figure 19.1, any excess demand for money at the interest rate r_1 bids up the interest rate to its equilibrium level r_0 . With excess demand for money, there is an excess supply of bonds. To make people want more bonds, suppliers of bonds offer a higher interest rate.³ People switch from money to bonds. The higher interest rate reduces both the excess supply of bonds and the excess demand for money. At the interest rate r_0 , money supply equals money demand. Bond supply equals bond demand. Both markets are in equilibrium. People wish to divide their wealth in precisely the ratio of the relative supplies of money and bonds.

From now on, we examine the implicit market for money. However, any statement about the money market is also a statement about the bond market.

Changes in equilibrium

A shift in either money supply or money demand changes equilibrium in the money market (and the bond market). These shifts are examined in Figure 19.2.

A fall in the money supply

Suppose the central bank lowers the money supply. For a fixed price level, lower nominal money reduces the real money supply. Figure 19.2 shows this leftward shift in the supply curve. Real money falls from L_0 to L_9 . The equilibrium interest rate rises from r_0 to r_9 . A higher interest rate reduces the demand for real money in line with the lower quantity supplied. Hence a lower real money supply raises the equilibrium interest rate. Conversely, a rise in the real money supply reduces the equilibrium interest rate.

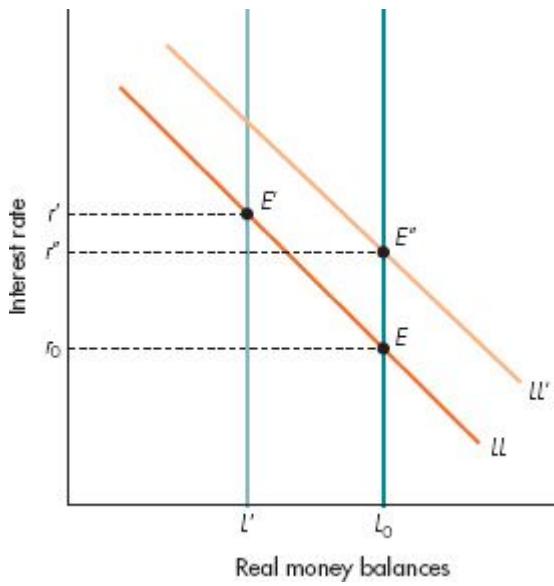


Figure 19.2 Equilibrium interest rates

A rise in real income

Figure 19.2 shows real money demand LL for a given real income. A rise in real income increases the marginal benefit of holding money at each interest rate, raising real money demand from LL to LL' . The equilibrium interest rate rises to keep real money demand equal to the unchanged real supply L_0 . Conversely, a fall in real income shifts LL to the left and reduces the equilibrium interest rate.

More competition in banking

Figure 19.2 also draws money demand LL for a given interest rate paid on bank deposits. Holding this rate constant, a rise in bond interest rates r raises the cost of holding money and reduces the quantity of money demanded. This implies the economy moves up a given demand curve LL .

However, more competition between banks, reflected in permanently higher interest rates paid on bank deposits, reduces the cost of holding money at each level of r . By raising money demand at each interest rate r , this shifts the demand for money up from LL to LL' . For a given money supply, this equilibrium interest rate on bonds is higher.

To sum up, a higher real money supply reduces the equilibrium interest rate, raising real money demand in line with the higher real money supply. Conversely, higher real income, which tends to raise real money demand, must lead to a rise in the equilibrium interest rate, which tends to reduce real money demand. Only then does real money demand remain equal to the unchanged supply. An increase in banking competition has similar effects to a rise in real income.

If attempts to regulate banks more closely, in order to reduce the risk of future financial crises, have the effect of *reducing* the competition between banks, then we can use the above analysis in reverse. The cost of holding money will increase, and the demand for money will shift downwards.

MATHS 19.1

SHIFTS IN MONEY DEMAND

Suppose the supply of real money is LS and that demand for real money is given by

$$LD = \alpha + \beta Y - \gamma(r - r_d) \quad \alpha > 0, \beta > 0, \gamma > 0$$

so that money demand rises with the level of income Y , but falls with the opportunity cost of holding money $(r - r_d)$, the difference between the interest rate r that could be earned by investing wealth elsewhere and the lower interest rate r_d that can be earned on bank deposits (often this is zero). In money market equilibrium, $LS = LD$. Changes in income and output Y move us along a *given* money demand curve, but changes in autonomous money demand α or in the cost of holding money $(r - r_d)$ lead to a shift in money demand.

For given levels of LS , Y and r_d , there is only one level of r that ensures

$$LS = LD = \alpha + \beta Y - \gamma(r - r_d)$$

An exogenous increase in money supply LS requires a fall in r to increase money demand in line with the new higher money supply. An exogenous increase in r_d , by reducing the opportunity cost of holding money, increases money demand. Yet, if money supply is unaltered, in equilibrium money demand cannot be allowed to

increase. This is achieved by an increase in the interest rate r on other assets, to restore the cost of holding money ($r - r_d$) to its original level. In this case, the increase in bank deposit rates is matched by an increase in the interest rate on other assets.

In all these examples, we can think of interest rates as adjusting almost instantaneously; that is, much more quickly than output and income adjust. It is therefore interest rate adjustments that ensure almost continuous equilibrium in the market for money.

19.5 Monetary control

The central bank can control the money supply by using open market operations to affect cash in circulation, or by using reserve requirements and the discount rate to affect the incentive of banks to create deposits, thereby affecting the money multiplier. This is easy in theory, but not so in practice.

It is hard for the Bank to control cash, because it is also lender of last resort. When the banks wish to increase lending and deposits they can always get extra cash from the Bank.

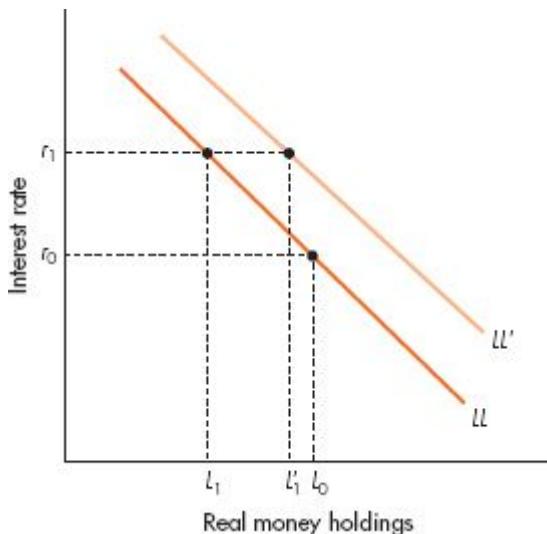
Nor, for any given quantity of cash, are deposits easily manipulated. To affect them, reserve requirements must force banks to hold reserves they would not otherwise have held. This is a tax on banks, stopping them conducting profitable business. Modern banks operating in global markets find ways around these controls. UK banks do business with UK borrowers using financial markets in Frankfurt or New York, and London is disadvantaged as a global financial centre.

Although the European Central Bank still retains minimum reserve ratios, the UK has given up required reserve ratios on banks for the purpose of monetary control. Since 2006, the Bank rewards banks for announcing their reserves at the Bank and sticking to them within the month. In essence, this allows the Bank to forecast the largest part of bank reserves. However, to translate this into a forecast for broad money, the Bank then has to forecast the size of the bank deposit multiplier. Concept 18.2 showed how volatile this can be, especially in a crisis when it really matters.

Hence precise control of broad money is difficult if viewed as the process of controlling narrow money or the money multiplier. Most central banks no longer try. Instead, they focus on broad money directly, and set interest rates to affect the demand for money, then passively supply whatever narrow money is necessary to bring broad money supply in line with this broad money demand. The TV news reports central bank decisions on interest rates, not its decisions on the money supply.

Control through interest rates

Figure 19.3 shows again the market for money. We draw the money demand schedule LL for a given level of real income. If the central bank can control the money supply, then, for a given level of goods prices, it can fix the real money supply at L_0 . The equilibrium interest rate is r_0 . Instead, the central bank can fix the interest rate at r_0 and supply the money needed to clear the market at this interest rate. In equilibrium, the central bank supplies L_0 .



The money demand schedule LL is drawn for a given level of real income. If the Bank can fix the real money supply at L_0 the equilibrium interest rate will be r_0 . Alternatively, if the Bank sets the interest rate r_0 and provides whatever money is demanded, the money supply will again be L_0 . To control the money supply by using interest rates, the Bank must know the position of the demand schedule. Fixing an interest rate r_1 , the resulting money supply will be L_1 if the demand schedule is LL but will be L'_1 if the demand schedule is LL' .

Figure 19.3 Interest rates and monetary control

The central bank can fix the money supply and accept the equilibrium interest rate implied by the money demand equation, or it can fix the interest rate and accept the equilibrium money supply implied by money demand. Central banks now do the latter.

Uncertainty about the exact size of the money multiplier or bank deposit multiplier is now unimportant. When the interest rate starts to fall below the level r_0 , either because of too little demand for money or too much supply, the Bank reduces

the monetary base, through an open market operation, until the interest rate is again r_0 . Conversely, when the interest rate exceeds r_0 , the Bank simply increases the monetary base until the interest rate falls to r_0 .

CONCEPT 19.1

QUANTITATIVE EASING

Quantitative easing is the creation of plentiful bank reserves to offset a lower bank deposit multiplier, preventing large falls in bank lending and broad money.

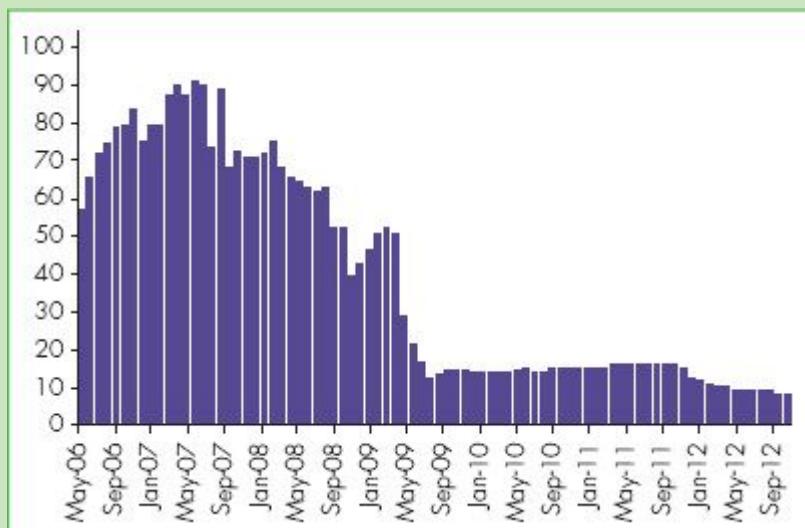
Banks responded to the financial crisis by prioritizing the rebuilding of their solvency. There were four aspects to this response: (a) holding a much higher percentage of their assets in ultra-safe bank reserves and other very liquid assets; (b) avoiding any new lending that was thought to be risky; (c) raising profit margins throughout the industry in order to build up capital reserves; and, where possible, (d) issuing new shares in order to attract additional capital from shareholders. Here we focus on the implications of (a) and (b).

In the charts below, we show again the collapse of the bank deposit multiplier – the ratio of broad money to bank reserves – which fell from 90 in mid-2007 to just 7 by 2012. If reserves had remained constant, broad money would have fallen to a thirteenth of its previous level! The complete drying up of bank lending – to each other and to private firms – transmitted a huge shock to the real economy. House prices fell since new mortgages were hard to get,

industrial production fell as firms struggled to finance work-in-progress until it could be sold, and increasing numbers of bankruptcies were reported.

The Treasury tried to help, by making it a condition of government support for banks such as RBS that they continued to lend to the private sector at the same level as in previous years. Unsurprisingly, the banks said they would but then failed to do so; there was little that the government could do.

UK and US central banks have usually been run by professional bankers, not world-class economics professors. At the time of the crisis, the governor of the Bank of England happened to be Mervyn King, former professor at the London School of Economics, and the governor of the US Federal Reserve was Ben Bernanke, former economics professor at Princeton University. They understood the problem and adopted a bold solution: quantitative easing (QE).



Bank deposit multiplier, UK

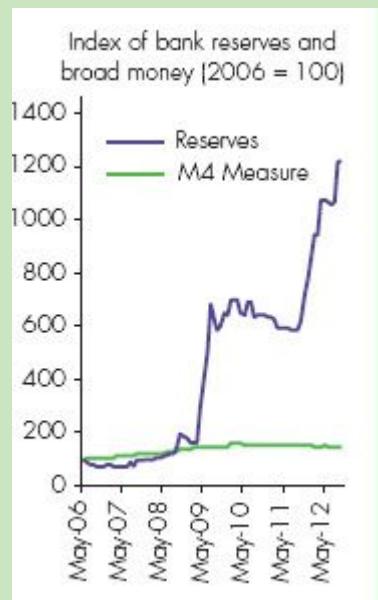
Source: Bank of England.

The first broad money and nominal GDP figure shows the evolution of the reserves of the UK banks and of the M4 measure of broad money, in each case using an index that sets the May 2006 level equal to 100. The QE programme began in the UK in 2009 and was extended in subsequent years. A similar programme was conducted in the US.

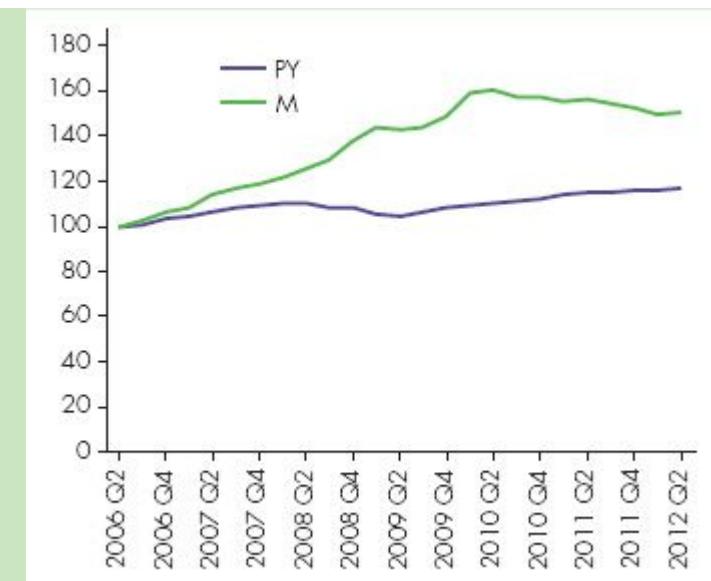
The reserves of UK banks rose six-fold between May 2008 and July 2009 as the Bank of England took action. The consequence was

achieving steady growth in broad money – notice from the second broad money figure that there was no spike in broad money growth. Broad money was nevertheless 60 per cent higher at the end of 2010 than it had been in mid-2006. This raises three obvious questions. (a) How did the Bank achieve quantitative easing? (b) Why did it want broad money to grow by 50 per cent when real output was stagnating? (c) Is inflation just around the corner?

To accomplish QE, the Bank announced that it would buy ‘safe’ bonds from private firms or government, in quantities that made this the mother of all open market operations. This put narrow money into the system. The charts show that most of this injection of narrow money, having circulated around the system a few times, ended up being held by banks as reserves at the Bank of England. Banks were still afraid of lending very much. But overall lending did increase. From May 2008 to July 2009, banks’ reserves increased from £27 billion to £152 billion, whereas broad money increased from £1737 billion to £2001 billion – so the £264 billion increase in broad money was caused not only by the £125 billion in bank reserves. As banks felt a little safer, they lent a little more, thereby raising bank deposits.



Broad money and nominal GDP



Broad money (M) and nominal GDP (PY), 2006 5 100

Source: www.bankofengland.co.uk/statistics.

The Bank of England's own research suggests that UK quantitative easing reduced the interest rate on long assets by up to 1 percentage point. Is this success or failure? The real interest rate matters for long-run decisions on investment by firms and consumption by households. Suppose nominal interest rates are

5 per cent and inflation is 3 per cent; the real interest rate is thus 2 per cent. If QE reduces the long nominal interest rate by 1 per cent, the real interest rate is now only 1 per cent, half its former level. This is a material cut in the real cost of borrowing. For firms and households that are optimistic about the future, it may stimulate additional spending. However, for pessimistic firms and households, expecting low demand in the immediate future, there is little need for extra capital equipment or consumer durables, however cheap it is to borrow. QE may be a useful boost to aggregate demand, but alone it cannot restore demand to previous levels.

Why was broad money allowed to grow so much despite the fact that the real economy was going backwards? The Bank of England was doing everything it could to stimulate economic recovery. Interest rates were reduced to near zero, which in itself raised the demand for money, which the Bank was then happy to see supplied. Thus, the main explanation of the rise in broad money is a sharp reduction in the cost of holding money. The second broad money

figure shows that increases in real output and the price level, and hence in nominal GDP, played little role in increasing the demand for broad money.

Although we defer our discussion of inflation until Chapter 22, we are already in a position to sketch an answer to our third question. If the economy is at full capacity, one might expect a 50 per cent increase in the broad money supply, and the interest rate reductions that presumably accompanied this, to cause a large rise in aggregate demand, well above the economy's capacity to supply – a recipe for a surge in inflation.

However, when the economy is facing its sharpest output downturn since the Great Depression, private firms and households are in no mood to spend. The immediate task is to raise aggregate demand back to acceptable levels. If and when that is accomplished, confidence will return. The proper task for the central bank is then to reverse the quantitative easing, reduce the money supply to more normal levels, and raise interest rates to the levels then required to prevent recovery spilling over into excess demand.

If it is technically possible to inject so much narrow money in such a short time, it is technically possible to do the reverse – the Bank sells the bonds it has recently acquired and receives narrow money in exchange, which is then 'retired' from circulation. Narrow money falls, and broad money falls even more as the normal bank deposit multiplier takes effect.

The key issue concerning financial markets is how the Treasury will then cope. During quantitative easing, it has been a simple matter to sell government debt to cover the budget deficit – if necessary, the Bank of England will buy it. Once the Bank is no longer a buyer but now an active seller of government debt, many private buyers must be found. This could cause a collapse in bond prices or, equivalently, a rise in the interest rates the government must pay to finance its debt.

If it became politically impossible for central banks to withdraw narrow money as confidence and the money multiplier increased, then the consequence would indeed be a surge in the broad money supply at that point, and this would almost certainly be inflationary.

In deciding to undertake QE on such a scale, central banks decided that these possible future outcomes were the lesser of two evils – without QE, the severe cutback in bank lending would have crippled the private sector in the immediate present. We shall have to see whether, at some future date, unwinding QE is as easy as initiating it.

19.6 Targets and instruments of monetary policy

Setting the interest rate not the money supply fineses the question of how the central bank forecasts the bank deposit multiplier. It also has a second advantage. When money demand is uncertain, fixing the money supply makes the interest rate uncertain; whereas fixing the interest rate makes the money supply uncertain. If the *effects* of monetary policy on the rest of the economy operate mainly via the interest rate, it is better to view monetary policy as the choice of interest rates not the money supply. In normal times, this is usually the case. However, if credit is in effect rationed because the banks are too scared to lend, the interest rate is not the whole story. Since we do not live in a permanent crisis – otherwise we would not call it a crisis – we revert to the discussion of monetary policy when the weather is less stormy.

Two other concepts guide our discussion of monetary policy in later chapters. One is the *ultimate objective* of monetary policy. Possible objectives could include price stability, output stabilization, influencing the exchange rate and reducing swings in house prices.

The **monetary instrument** is the variable over which the central bank makes day-to-day choices.

An **intermediate target** is a key indicator used to guide interest rate decisions.

To pursue its ultimate objective, what information does a central bank use at its frequent meetings to decide interest rates? It gets up-to-date forecasts of many variables. Sometimes, it concentrates on one or two key indicators.

Interest rates are the **instrument** about which policy decisions are made, but interest rates are chosen to try to keep the **intermediate target** on track.

This shows how interest rates should adjust to the state of the economy. New data on the money supply (largely bank deposits) come out faster than new data on the price level or output. In the heyday of monetarism, central banks changed interest rates to try to meet medium-run targets for the path of nominal growth. In terms of Figure 19.3, it was as if they were fixing the money supply, not interest rates.

Throughout the world, in the past two decades there have been two key changes in the design of monetary policy. First, central banks have been told that their ultimate objectives should concentrate more on price stability.

Second, money has become less important as an intermediate target. The financial revolution reduced its reliability as a leading indicator of future inflation. When structural changes in the financial sector are causing changes in money demand, it is hard to predict how much money will be held and how much will be spent. Increasingly, central banks use *inflation targets* as the intermediate target to which interest rate policy responds.

MATHS 19.2

QUANTITATIVE EASING REVISITED

Suppose money demand is given by

$$M = aY - br \quad a > 0, b > 0 \quad (1)$$

where r is the nominal interest rate. Money demand is higher if income is higher (greater benefit to holding money), but lower if interest rates are higher (higher cost of holding money). Broad money is related to bank reserves R via the bank deposit multiplier m :

$$M = mR \quad m > 0 \quad (2)$$

Aggregate demand AD increases with autonomous demand A but is reduced by higher interest rates:

$$AD = A - hr \quad A > 0, h > 0 \quad (3)$$

Consider a fall in autonomous demand from A to lA , where $0 < l < 1$. Simultaneously, banks get scared and the bank deposit multiplier falls from m to ρm , where $0 < \rho < 1$. By how much does the central bank need to increase bank reserves R in order to maintain aggregate demand for output at its original level? Originally,

$$mR = M = aY - br \quad (4)$$

However, $r = (A - AD)/h = (A - Y)/h$, once short-run equilibrium equates output and aggregate demand, hence substituting this value of r into equation (4):

$$R = [aY - b(A - Y)/h]/m = [(ah + b)Y - bA]/hm \quad (5)$$

For a given level of autonomous demand A , aggregate demand and output can be higher only if interest rates are lower, which requires a larger money supply, for which a larger quantity of narrow money is necessary if there is a fixed ratio of broad money to narrow money. Conversely, if autonomous demand A is higher, output and aggregate demand can remain fixed only if interest rates increase to offset the rise in autonomous demand, for which a reduction in the money supply is necessary. Hence, with a fixed bank deposit multiplier, reserves must be lower.

When the deposit multiplier falls from m to ρm , this effect alone induces the central bank to increase R in order to maintain output at its former level. The new level of reserves would be R/ρ . R , since $\rho < 1$. Additionally, the fall in autonomous demand from A to lA raises the level of reserves that the central bank must supply if output and aggregate demand are not to fall. Quantitative easing is the central bank response to twin problems, with the same cause: lower aggregate demand and a lower bank deposit multiplier.

19.7

The transmission mechanism

The central bank sets interest rates. How do interest rates affect the real economy?

In a closed economy, monetary policy affects consumption and investment demand by affecting real interest rates.⁴ The central bank chooses the nominal interest rate. If prices are fixed, this is also the real interest rate. Once we allow prices to vary, monetary policy needs to anticipate what inflation will be. Since the real interest rate is simply the nominal interest rate minus the inflation rate, monetary policy then sets the nominal interest rate to get the desired real interest rate.

The **transmission mechanism** of monetary policy is the channel through which it affects output and employment.

Consumption demand revisited

Chapter 16 used a very simple consumption function, an upward-sloping straight line relating aggregate consumption to the disposable income of households. The slope of this line, the marginal propensity to consume, showed the fraction of each extra pound of disposable income that households wished to spend, not save.

The height of the consumption function showed autonomous consumption demand; that is, the part unrelated to personal disposable income. Changes in disposable income moved households *along* the consumption function. Changes in autonomous demand *shifted* the consumption function. How can monetary policy affect autonomous consumption demand?

Household wealth

The **wealth effect** is the shift in the consumption function when household wealth changes.

Suppose real **wealth** rises because of a stock market boom. Households spend some of their extra wealth on a new car. At each level of disposable income, consumption demand is higher. The entire consumption function shifts up when household wealth increases.

Money and interest rates affect household wealth, and thus consumption and aggregate demand, in two ways. First, since money is a component of household wealth, a higher real money supply adds directly to

household wealth. Second, interest rates affect household wealth indirectly. The price of company shares and long-term government bonds is the present value of the expected stream of dividend earnings or promised coupon payments. When interest rates fall, future earnings, now discounted at a lower interest rate, are worth more today. Lower interest rates make the price of bonds and corporate shares rise and make households wealthier.⁵

Durables and consumer credit

When spending exceeds disposable income, net wealth falls. People sell off assets or borrow money to finance their dissaving. A lot of borrowing is to finance purchases of *consumer durables*; that is, household capital goods such as televisions, furniture and cars. Splashing out on a new car can cost a whole year's income.

Two aspects of consumer credit or borrowing possibilities affect consumption spending. First, there is the quantity of credit on offer. If banks or retailers make more credit available to customers, people are more likely to buy the car or dream kitchen they have always wanted. An increase in the supply of consumer credit shifts the consumption function upwards. People spend more at any level of disposable income. Second, the cost of credit matters. The higher the interest rate, the lower the quantity that households can borrow while still being able to make repayments out of their future disposable incomes.

Money and interest rates thus affect consumer spending by affecting both the quantity of consumer credit and the interest rates charged on it. An increase in the monetary base increases the cash reserves of the banking system and allows it to extend more consumer credit in the form of overdrafts. And by reducing the cost of consumer credit, lower interest rates allow households to take out bigger loans while still being able to meet the interest and repayments.

Those two forces – wealth effects and changes in consumer credit – explain most of the shifts in the consumption function. They are part of the *transmission mechanism* through which monetary policy affects output and employment. Operating through wealth effects or the supply and cost of consumer credit, changes in the money supply and in interest rates shift the consumption function and the aggregate demand schedule, thus affecting equilibrium income and output.

Two closely related theories of the consumption function reinterpret these phenomena and make some of their subtleties more explicit.

The permanent income hypothesis

Developed by Professor Milton Friedman, this hypothesis assumes that people's incomes fluctuate but that people dislike fluctuating consumption. Because of diminishing marginal utility, a few extra bottles of champagne in the good years does not compensate for hunger in the bad years. Rather than allow fluctuations in income to induce fluctuations in consumption, people smooth out fluctuations in consumption. People go without champagne to avoid being hungry.

The **permanent income hypothesis** says that consumption reflects long-run or permanent income.

What determines the consumption people can afford on average? Friedman coined the term **permanent income** to describe people's average income in the long run, and argued that consumption depends not on current disposable income but on permanent income.

Suppose people think current income is unusually high. This temporarily high income makes little difference to their permanent income or the consumption they can afford in the long run. Since permanent income has hardly risen, they hardly increase current consumption.

They save most of their temporary extra income and put money aside to see them through the years when income is unusually low. Only if people believe that a rise in today's income will be sustained as higher future incomes will their permanent income rise significantly. Only then is a large rise in current income matched by a large rise in current consumption.

The **life-cycle hypothesis** asserts that people make a lifetime consumption plan (including bequests to their children) just affordable out of lifetime income (plus any initial wealth inherited).

The life-cycle hypothesis

Developed by Professors Franco Modigliani and Albert Ando, this theory takes a long-run approach like the permanent income hypothesis, but recognizes that changing tastes over a lifetime may undermine complete consumption smoothing.

Each individual household need not plan a constant consumption level over its lifetime. There may be years of heavy expenditure (a round-the-world cruise, sending the children to private school) and other years when spending is a bit less. However, such individual discrepancies tend to cancel out in the aggregate. Like the permanent income hypothesis, the life-cycle hypothesis suggests that it is average long-run income that determines the total demand for consumer spending.

Figure 19.4 shows a household's actual income over its lifetime. Income rises with career seniority until

retirement, then drops to the lower level provided by a pension. The household's permanent income is OD . Technically, this is the constant annual income with the same present value as the present value of the actual stream of income. If the household consumed exactly its permanent income, it would consume OD each year and die penniless. The two areas labelled A show when the household would be spending more than its current income and the area B shows when the household would be saving.

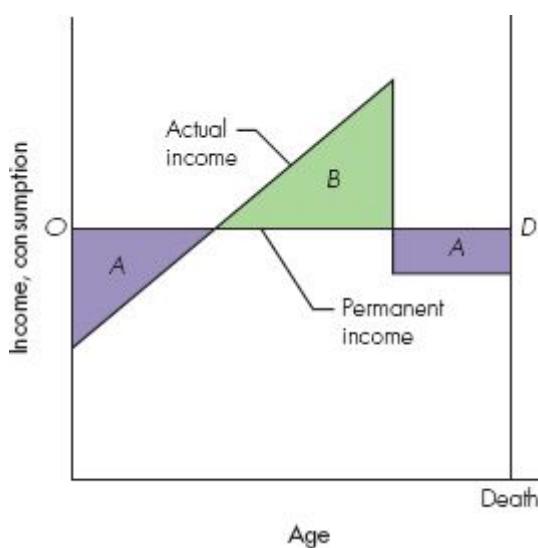
The household spends its income over its lifetime, but area B is not the sum of the two areas A because of compound interest. In the early years of low income, the household borrows. The area B shows how much the household has to save to pay back the initial borrowing *with interest* and accumulate sufficient wealth to see it through the final years when it is again dissaving.

Now let's think again about wealth effects and consumer credit. With more initial wealth, a household can spend more in every year of its lifetime without going broke. We can shift the permanent income line in Figure 19.4 upwards and consumption will rise. Although area B is now smaller and the areas A are now larger, the household can use its extra wealth to meet this shortfall between the years of saving (the area B) and the years of dissaving (the two areas A).

Again, we conclude that higher wealth leads to more consumption at any current disposable income, but we pick up something we missed earlier. If households believe their *future* income will be higher than previously

imagined, this also raises their permanent income. Households can spend more each year and still expect to balance their lifetime budget. They raise *current* consumption as soon as they raise their estimates of future incomes. The present value of future income plays a role very similar to wealth. It is money to be shared out in consumption over the lifetime. Friedman called it ‘human wealth’, to distinguish it from financial and physical assets. Rises in expected future incomes have wealth effects. They shift up the simple consumption function relating *current* consumption to *current* disposable income.

What about consumer credit? A rise in interest rates reduces the present value of future incomes and makes households worse off. In Figure 19.4, households must enlarge area *B* to meet the extra interest costs of paying back money borrowed in area *A* early in the lifetime. We must shift the permanent income line downwards. A rise in interest rates reduces current consumption not merely by reducing the market value of financial assets, but also by reducing the present value of future *labour* income. By reducing human wealth, it shifts the consumption function downwards.



Actual disposable income rises over a household's lifetime until retirement, then falls to the pension level. Permanent income is the constant income level OD with the same present value as actual income. Suppose consumption equals permanent income. The two areas A show total dissaving and the area B, total saving. In the absence of inherited wealth and bequests, B must be large enough to repay borrowing with interest and also build up enough wealth to supplement actual income during retirement.

Figure 19.4 Consumption and the life-cycle

Short-term interest rates apply to loans of very short maturity.

Long-term interest rates apply to long-term loans during which the interest rate is usually fixed.

From short-term interest rates to long-term interest rates

The market for money determines **short-term interest** rates, for loans of duration ranging from a day to a month. However, long-term investment decisions about building a factory that will last for 20 years need to consider **long-term interest rates** over a similar duration.

Sometimes, a 20-year loan will be at variable interest rates, as with a UK mortgage. In such a contract, the lender commits to lend for the long duration but varies the interest rate during the loan, in line with what is happening to short-term loans in the money market. Most other long-term loans (including mortgages in most continental European countries) entail a fixed rate of interest over the entire life of the contract.

Since firms could in principle finance long-term investment by a succession of short-term loans, there ought to be a close relation between the long-term interest rate today and the sequence of expected short-term interest rates over the same period.

Either way, the effect of interest rates on investment demand for long-term capital assets should not depend solely on today's short-term interest rate but on the average of expected short-term interest rates during the life of the loan, plus a little extra for risk to compensate being locked into the long-term loan.

Governments therefore affect long-term decisions – for investment by firms and for some important longer-term spending by households – by affecting beliefs about current and future short-term interest rates. If changing the current interest rate has little effect on the long-term interest rate, then the transmission mechanism from monetary policy to aggregate demand will at best be weak.

ACTIVITY 19.1

TRANSMISSION LAG

In the UK, people rarely get access to fixed-rate mortgages for the entire life of the loan, but are sometimes offered interest rates fixed for the first two years. During this initial period, changes in short-term interest rates have little effect on such households, since the interest rate they care about most is fixed. However, as this initial period expires, households then have to face whatever the new level of interest rate has become.

Thus, UK households cannot fix their mortgage interest rate for an extended period. Taking a 20-year view, they basically have a variable-rate mortgage, with successive small steps in which the interest rate may be temporarily fixed.

In fact, the UK is quite unusual in having such a high proportion of variable-rate mortgages – many continental European countries fix the mortgage interest rate for 20 years or whatever the duration of the mortgage loan. This has important consequences for the transmission mechanism of monetary policy. When most households are immune to the mortgage impact of interest rate changes, then, other things equal, the central bank has to move the interest rate by more in order to have the same effect on aggregate demand. If interest rates do not work through existing mortgages, they have to work more on other determinants of aggregate demand. Conversely, since the UK is so exposed to the effect of interest rate changes in the short run, the Bank of England has a more powerful weapon with which to manage aggregate demand. Since the weapon works more effectively, it requires smaller interest rate changes to achieve the same effect.

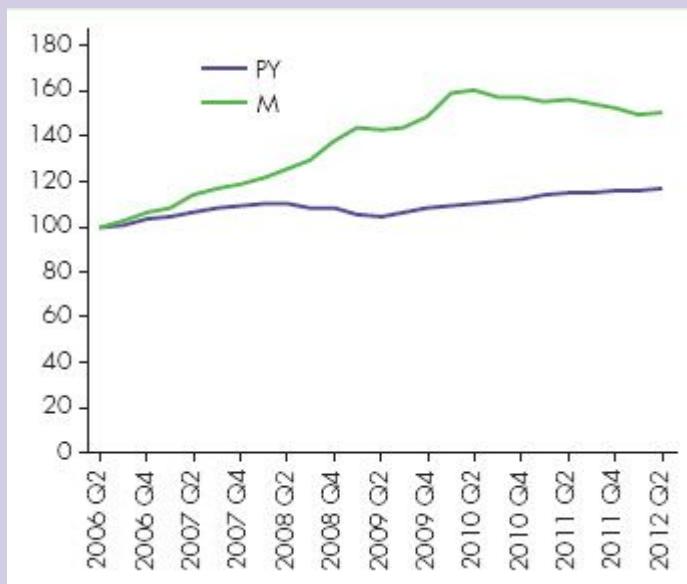
This is one of many reasons why changes in interest rates may take several years to have their full effect on aggregate demand. Overnight, people are locked into old contracts that shield them for a while from the effect of the new interest rate. Even after they feel its effect, it may take time to assess how painful it is and to look for alternative ways to behave.

In the US, Ben Bernanke, president of the US central bank, recognized in 2012 that, with US interest rates already near zero, monetary policy had lost its power to boost aggregate demand through yet lower short-term interest rates in 2012. However, he understood the relationship between long-term interest rates and short-term interest rates. So he promised that the US central bank

would keep interest rates near zero for several years to come. The purpose of this promise was not to change the short-term interest rate in 2012, but to reduce the long-term interest rate – relevant to aggregate demand – by changing beliefs about the short-term interest rates in the future that should be averaged to get the *long-term* interest rate in 2012.

Bernanke adopted other unconventional means to overcome the breakdown of the normal transmission mechanism of monetary policy. In Operation Twist, the US Federal Reserve bought massive quantities of long-term government bonds, inducing higher prices and lower interest rates on long bonds, financing these purchases not by printing money but by selling off the Fed's holdings of short bonds.

Bernanke also pursued quantitative easing – the purchase of large quantities of bonds by printing money – for the purpose of getting long-term interest rates down in order to boost aggregate demand. As in the UK, it is acknowledged that this policy should be unwound – selling bonds and removing cash from the market – once recovery is sustainably established.



Summary: the transmission mechanism of monetary policy

In the Eurozone, Central Bank President Mario Draghi became worried about a different transmission failure. Although all Eurozone countries face the same short-term interest rate – the one set by the European Central Bank (ECB) – different member states

faced very different long-term interest rates. In countries like Germany, long-term interest rates were low, foreseeing a long period of low short-term interest rates.

However, in Greece, Spain, Portugal, Italy and Ireland, long-term interest rates had become huge because of the fear of default on long-term government bonds. Draghi argued that this was preventing the ‘normal transmission mechanism of monetary policy’ in these countries. He promised that the ECB would buy these long-term bonds – in whatever quantities needed – to restore more normal relationships between short- and long-term interest rates, thereby allowing monetary policy to start working again.

Critics argued this was simply covert support for the budget deficits in these countries. Supporters argued that Draghi had saved the Eurozone by preventing peripheral countries being forced into bankruptcy by self-fulfilling prophecies that made the interest burden of their debt unsustainable. The figure below summarizes our discussion of the transmission mechanism of monetary policy.

Questions

- (a) Suppose we are creatures of habit – calculating optimal behaviour takes time and effort so we recalculate only rarely, when it has become obvious to us that circumstances have changed substantially and previous behaviour cannot possibly be optimal. Could this explain a delay in the transmission mechanism of monetary policy even if there are no long-term contracts in force? Give an example.
- (b) Would this justify a transmission lag in responses to fiscal policy too?
- (c) Suppose interest rates can be changed frequently whereas fiscal policy changes are infrequent. Would this help explain why people are slower to respond to monetary changes than fiscal changes?
- (d) Milton Friedman argued that the lags in monetary policy are ‘long and variable’. Using the figure above, outline the various points in the transmission mechanism at which this could apply.

To check your answers to these questions, go to page 680.

Finally, what about a rise in the quantity of consumer credit on offer? Figure 19.4 assumes that people spend more than their incomes early in life. Students run up overdrafts knowing that, as rich economists, they can pay them back later. What if nobody will lend? People without wealth are restricted by their actual incomes, although people with wealth can lend to themselves by running down their wealth. Hence a rise in the availability of consumer credit lets people dissave in the early years. Total consumption rises. More students run up overdrafts and buy cars.

Conversely, in the aftermath of a financial crash, banks are preoccupied with restoring their own solvency. They may be very reluctant to lend to any borrower who looks at all risky. In such circumstances, the supply of credit can contract sharply, forcing households to reduce their spending.

Having discussed how monetary policy affects consumption demand, we conclude our examination of monetary transmission by analysing how interest rates affect investment demand.

Investment demand

In earlier chapters we treated investment demand as autonomous, or independent of current income and output. We now begin to analyse what determines investment demand. Here we focus on interest rates.

Total investment spending is investment in fixed capital and investment in working capital. Fixed capital includes factories, houses, plant and machinery. The share of investment in GDP fluctuates between 10 and 20 per cent.⁶ Although the total change in inventories is quite small, this component of total investment is volatile and contributes significantly to changes in the total level of investment.

In a closed economy, aggregate demand is $C + I + G$. Public investment is part of G . We still treat government demand as part of fiscal policy. Thus we assume that G is fixed at a level set by the government. In this section we focus on private investment demand I .

Investment in fixed capital

Firms add to plant and equipment because they foresee profitable opportunities to expand output or because they can reduce costs by using more capital-intensive production methods. BT needs new equipment

because it is developing new products for data transmission. Nissan needs new assembly lines to substitute robots for workers in car production.

The firm weighs the benefits from new capital – the rise in profits – against the cost of investment. The benefit occurs in the future, but the costs are incurred when the plant is built or the machine bought. The firm compares the value of extra future profits with the current cost of the investment.

Will the investment yield enough extra profit to pay back *with interest* the loan used to finance the original investment? Equivalently, if the project is funded out of existing profits, will the new investment yield a return at least as great as the return that could have been earned by lending the money instead? The higher the interest rate, the larger must be the return on a new investment to match the opportunity cost of the funds tied up.

At any moment, there are many investment projects a firm *could* undertake. The firm ranks these projects, from the most profitable to the least profitable. At a high interest rate, only a few projects earn enough to cover the opportunity cost of the funds employed. As the interest rate falls, more and more projects earn a return at least matching the opportunity cost of the funds used to undertake the investment. The firm invests more.

The **investment demand schedule** shows the desired investment at each interest rate.

Figure 19.5 plots the **investment demand schedule** relating interest rates and investment demand.

If the interest rate rises from r_0 to r_1 , fewer investment projects cover the opportunity cost of the funds tied up, and desired investment falls from I_0 to I_1 . The height of the schedule II reflects the cost of new capital and the stream of profits to which it gives rise. For a given stream of expected future profits, a higher price of new capital goods reduces the return on the money tied up in investment. Fewer projects match the opportunity cost of any particular interest rate. Since desired investment

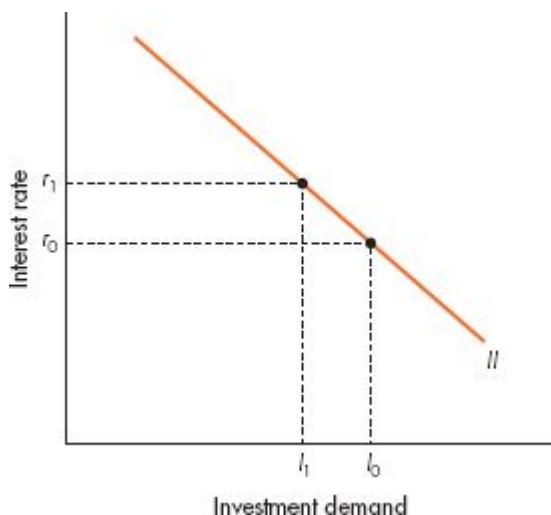
is then lower at any interest rate, a rise in the cost of new capital goods shifts the investment demand schedule II downwards.

Similarly, pessimism about future output demand reduces estimates of the stream of profits earned on possible investment projects. The return on each project falls. At each interest rate, fewer projects match the opportunity cost of the funds. Desired investment falls at any interest rate. Lower expected future demand shifts the investment demand schedule downwards.⁷

The investment demand schedule II can be used to analyse both business investment in plant and machinery and residential investment in housing. What about the slope of the schedule? There is a big difference between a machine that wears out in three years and a house or a factory lasting 50 years. The longer the economic life of the capital good, the larger the fraction of its total returns earned in the distant future, and the more the original cost of the goods accumulates at compound interest before the money is repaid.

Hence a change in short-term interest rates today has two effects. First, it affects the long-term interest rates relevant to the life of a loan to finance a long-term capital good. Second, for a given change in long-term interest rates, the investment demand schedule is flatter, and the monetary transmission mechanism more powerful, for long-lived houses and factories than for short-term machinery.⁸ A change in interest rates has more effect on long-term projects.

This conclusion has to be qualified with the caveat that a temporary change in short-term interest rates, achieved through monetary policy, may have little effect on long-term interest rates if it is believed today that future short-term interest rates will be largely unaffected. However, a change in short-term interest rates that is believed today to be permanent will have a powerful effect on long-term rates and hence on long-term investment.



For a given price of capital goods and given expectations about the profit stream to which new investments give rise, a higher interest rate reduces the number of projects that can provide a return matching the opportunity cost of the funds used. As interest rates rise from r_0 to r_1 , desired investment falls from I_0 to I_1 .

Figure 19.5 The investment demand schedule

Inventory investment

There are three reasons why firms desire stocks of raw materials, partly finished goods and finished goods awaiting sale. First, the firm may be betting on price changes. Sometimes, firms hold large stocks of oil, believing it cheaper to buy now rather than later. Similarly, firms may hold finished goods off the market hoping to get a better price later.

Second, many production processes take time. A ship cannot be built in a month, or even a year. Some stocks are simply the throughput of inputs on their way to becoming outputs.

Third, stocks help smooth costly adjustments in output. If output demand rises suddenly, plant capacity cannot be changed overnight. A firm has to pay big overtime payments to meet the upsurge in orders. It is cheaper to carry some stocks, available to meet a sudden rise in demand. Similarly, in a temporary downturn, it is cheaper to maintain output and pile up stocks of unsold goods than to incur expensive redundancy payments to cut the workforce and reduce production.

CONCEPT 19.2

THE CREDIT CHANNEL OF MONETARY POLICY

The **credit channel** affects the value of collateral for loans, and thus the supply of credit.

Recent research emphasizes that interest rates are not the only **channel** through which monetary policy affects consumption and investment, and hence aggregate demand.

A lender usually asks for collateral – assets available for sale if you fail to repay the loan. Collateral is how lenders cope with moral hazard and adverse selection: borrowers who know more about their ability and willingness to repay than lenders know.

Suppose the price of goods falls, raising the real value of nominal assets. People have more collateral to offer lenders, who lend more than before at any particular interest rate. The supply of credit rises, and aggregate demand for goods increases.

There are really two credit channels, since there are two reasons for changes in the value of collateral. First, changes in goods prices change the real value of nominal assets. Second, and quite distinct, when monetary policy changes the interest rate, this affects the present value of future income from assets and the market value of collateral assets themselves.

This theoretical reasoning is supported by evidence from quantitative easing. Its purpose was not simply to raise the broad money supply to support the desired low level of interest rates. It was also believed that credit rationing by lenders was curtailing private spending. Injecting more money provided additional liquidity to people who would otherwise have been credit-rationed, and the consequent spending helped bid up house prices and share prices on the stock market. In turn, this improved private sector collateral and made banks more willing to lend, causing a second-round beneficial effect.

In this extreme example, raising the money supply has beneficial effects despite the fact that interest rates have already fallen as low as they can go.

John Maynard Keynes thought monetary policy became powerless once interest rates had been driven down to zero, what he called the *liquidity trap*. Great man that he was, he did not get everything right. Nowadays, we know about the credit channel, and quantitative easing is proof that it can work.

These are benefits of holding inventories. The cost is that, by retaining unsold goods or buying goods not yet inputs to production, a firm ties up money that could have earned interest. The cost of holding inventories is the interest forgone, plus any storage charges for holding stocks.

Thus the investment demand schedule *II* for fixed capital in Figure 19.5 also applies to increases in working capital, or inventories. Other things equal, a higher interest rate reduces desired stockbuilding, an upward move *along* the investment demand schedule. This is part of the monetary transmission mechanism. But a rise in potential speculative profits, or fall in storage costs for inventories, *shifts* the schedule *II* up and raises inventory investment at any interest rate. Not all changes in investment demand are caused by monetary policy.

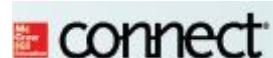
Summary

- The Bank of England, the UK **central bank**, is banker to the banks. Because it can print money it can never go bust. It acts as **lender of last resort** to the banks.
- The Bank conducts the government's monetary policy. It affects the monetary base through **open market operations**, buying and selling government securities. It can also affect the money multiplier by imposing **reserve requirements** on the banks, or by setting the **discount rate** for loans to banks at a penalty level that encourages banks to hold excess reserves.

- There is no explicit market in money. Because people plan to hold the total supply of assets that they own, any excess supply of bonds is matched by an excess demand for money. Interest rates adjust to clear the market for bonds. In so doing, they clear the money market.
- A rise in the real money supply reduces the equilibrium interest rate. For a given real money supply, a rise in real income raises the equilibrium interest rate.
- In practice, the Bank cannot control the money supply exactly. Imposing artificial regulations drives banking business into unregulated channels. **Monetary base control** is difficult since the Bank acts as lender of last resort, supplying cash when banks need it.
- Thus the Bank sets the interest rate not money supply. The demand for money at this interest rate determines the quantity of money supplied. **Interest rates are the instrument of monetary policy.**
- Interest rates take time to affect the economy. **Intermediate targets** are used as leading indicators when setting the interest rate.
- **Quantitative easing** is the creation of substantial quantities of bank reserves in order to offset a fall in the bank deposit multiplier and prevent large falls in bank lending and broad money.
- A higher interest rate reduces household wealth and makes borrowing dearer. Together, these effects reduce autonomous consumption demand and shift the consumption function downwards.
- **Consumption demand** reflects long-run disposable income and a desire to smooth out short-run fluctuations in consumption. Higher interest rates reduce consumption demand by reducing the present value of expected future labour income.
- Given the cost of new capital goods and expected stream of future profits, a higher interest rate reduces **investment demand**, a movement down a given investment demand schedule *II*. Higher expected future profits, or cheaper capital goods, shift the *II* schedule upwards.

- Long-term interest rates are an average of the sequence of short-term interest rates expected to prevail during the life of the long loan contract, plus an extra allowance for risk.
- The effect of change in short-term interest rates on consumption and investment demand, often operating through an induced change in the long-term interest rate, is the **transmission mechanism** of monetary policy.

Review questions



EASY

- 1 The Bank of England sells £1 million of securities to Mr Jones who banks with Barclays. The Bank of England requires other commercial banks to hold 100 per cent cash reserves against deposits. What is the money multiplier?
- 2 Which of these statements is correct? The purpose of quantitative easing is: (a) to create money in order to create inflation and reduce the real value of government debt; (b) to force banks to create deposits despite having inadequate bank reserves; (c) to make the central bank the purchaser of last resort for government bond issues; (d) to prevent a collapse of broad money when banks are unable or unwilling to lend.
- 3 What are the desirable properties of a good leading indicator for interest rate decisions?
- 4 What happens to the consumption function if the banks decide to offer credit cards on easier terms and conditions? Why?
- 5 **Common fallacies** Why are these statements wrong? (a) By abolishing reserve requirements, the central bank gave up any attempt to control the money supply. (b) When real interest rates are negative, people are being paid to hold cash. (c) Consumers are said to behave irrationally if their spending is up when their disposable income is lower.

MEDIUM

- 6 The Bank of England sells £1 million of securities to Mr Jones who banks with Barclays. (a) If Mr Jones pays by cheque, show the effect on the balance sheets of the Bank of England and Barclays. (b) What

- happens to the money supply? (c) Is the answer the same if Mr Jones pays in cash?
- 7 Using a diagram like Figure 19.1: (a) illustrate the initial effects of a recovery in confidence after a financial crash. (b) Did you expect a larger effect on money supply or money demand? (c) How will monetary policy then respond?
- 8 Suppose banks begin lending again as confidence is restored. (a) If monetary policy takes no action, what will be the likely outcome? (b) What action by the central bank would then be appropriate?
- 9 Why might it take up to two years for a change in interest rates fully to affect aggregate demand? What does this imply about decisions to set interest rates?
- 10 If the permanent income hypothesis is correct, we should expect to see a lower marginal propensity to consume in the short run than in the long run. Why?
- 11 Why do higher interest rates reduce investment demand? Be sure to discuss all the different ways in which firms might finance their investment projects.

HARD

- 12 Consider a simplified version of the model in Maths 19.2, in which money demand, the deposit multiplier and aggregate demand are, respectively: $M = Y - r$ $M = mR$ $Y = 100 - r$
- If output is 90, find the interest rate and the level of bank reserves and the level of bank reserves.
 - suppose, autonomous aggregate demand falls from 100 to 95, and the deposit multiplier falls from m to $m/2$. How much must reserves increase to preserve the initial level of output? (c) What is the maximum fall in autonomous aggregate demand that can be offset by quantitative easing?
- 13 **Essay question** Why do modern central banks think of monetary policy as choosing the interest rate rather than the money supply?
- 14 You live for five periods, during which, respectively, you earn 100, 200, 300, 200 and 100. (a) Draw a diagram of your lifecycle income. (b) If the interest rate is zero, and there is no inflation, what is your permanent income? (c) Using your diagram, or otherwise, identify your saving or dissaving in each period of your life. (d) If the real interest rate is positive instead of zero, what effect does this have on your initial estimate of your permanent income? Illustrate in your diagram.

- 1 Financial regulation is sometimes the responsibility of the central bank, but sometimes the responsibility of a separate financial regulator. In the UK, responsibility was transferred from the Bank of England to the Financial Services Agency in 1997.
- 3 A bond is a promise to pay a given stream of interest payments over a given time period. The bond price is the present value of this stream of payments. The higher the interest rate at which the stream is discounted, the lower the price of a bond. With an excess supply of bonds, bond prices fall and the interest rate or rate of return on bonds rises.
- 4 In Chapter 29 we show that, in an open economy, there is also a strong relationship between interest rates, the exchange rate and competitiveness. Monetary transmission then includes effects on export and import demand.
- 5 When interest rates are 10 per cent, a bond paying £2.50 for ever is worth £25. New buyers get about 10 per cent a year on their investment. If interest rates fall to 5 per cent, bond prices rise to £50. New buyers still get an annual return in line with interest rates on other assets. A similar argument applies to company shares.
- 6 These numbers refer to gross investment: the production of new capital goods that contribute to aggregate demand. Since the capital stock is depreciating, or wearing out, some gross investment is needed merely to keep the existing capital stock from falling.
- 7 We can make the same points another way. Given the stream of future profits and the interest rate, a firm does all projects for which the present value of operating profits exceeds the initial price of the capital goods. A higher interest rate cuts the present value of profits. Some projects no longer cover the initial cost of capital goods. Higher interest rates reduce desired investment. Similarly, a lower expected future profit stream, or higher purchase price of capital goods, cuts the present value of operating profits relative to the initial cost, reducing investment demand.
- 8 Equivalently, a 1 per cent rise in the interest rate has a small effect on the present value of earnings over a three-year period but a large effect on the present value of earnings over the next 50 years. Note that this is the same argument as we used in Chapter 17, in saying that a change in interest rates would have little effect on the price (present value of promised payments) of a short-term bond but a large effect on the price of a long-term bond.

CHAPTER 20

Monetary and fiscal policy

Learning outcomes

By the end of this chapter, you should be able to:

- 1 describe different forms of monetary policy
- 2 recognise a monetary target
- 3 understand the *IS* and *LM* schedules
- 4 understand equilibrium in both the output and money markets
- 5 describe the effect of a fiscal expansion
- 6 describe the effect of a monetary expansion
- 7 understand the mix of monetary and fiscal policy
- 8 realise how expected future taxes affect current demand

Chapters 16 and 17 introduced a simple model of income determination, and studied how fiscal policy affects aggregate demand and equilibrium output. Chapters 18 and 19 examined the demand for money, the supply of money and the determination of interest rates. Interest rates connect the present and the future, affecting spending decisions of both households and firms. We analysed the transmission mechanism by which monetary policy affects aggregate demand.

We now examine the interaction of the markets for goods and for money. Interest rates affect the demand for goods and the level of income and output, but income and output affect the demand for money and the interest rates set by the central bank.

We need to think about both markets at once. In so doing, we explain how equilibrium income and interest rates are simultaneously determined. In this richer model, we study changes in monetary and

fiscal policy. Finally, we discuss how the mix of monetary and fiscal policy affects the composition as well as the level of equilibrium output.

This is the last chapter in which we retain the simplifying assumption that prices are fixed. The interest rate is the key variable connecting the markets for money and output. In the next chapter, we allow prices to change, and introduce aggregate supply for the first time.

20.1 Monetary policy

Economists distinguish between rules and discretion. A smoker decides from minute to minute whether to have a cigarette. Once we understand her preferences, the price of cigarettes, her income and the attitude of her friends, we can model her behaviour and predict how she will behave, even though she has discretion or freedom to decide how much to smoke.

A **rule** is a commitment on how to behave, for example to smoke no more than ten cigarettes a day. This rule is credible only if we understand what prevents the smoker having the eleventh cigarette when she desperately wants one. Losing her friends if she smokes more than ten might be a commitment mechanism to enforce the rule. A rule constrains her **discretion**, limits her freedom and precludes the choice she would otherwise have made.

A **rule** is a commitment describing how behaviour changes as circumstances change. **Discretion** means free choice without restrictions imposed by prior commitments.

A particular **monetary policy** is a relationship between the state of the economy and the interest rate chosen by the central bank.

What do we mean by a given **monetary policy**? This has two aspects. First, to what variable does it refer – the interest rate or the money supply? For the reasons given in the previous two chapters, we prefer to focus on the interest rate.

Second, does a given policy mean the choice of a particular interest rate? Changing the interest rate would then be a change in policy. This is simple, but we can do better. We can usually model *why* that interest rate was chosen: the relationship between the chosen interest rate and other economic variables.

Interest rates change either because economic circumstances change (within a *given* monetary policy) or because the central bank switches to a different preferred relationship between interest rates and the state of the economy (a *change* in monetary policy).

In the heyday of monetarism, central banks used to adjust interest rates to stop the money supply deviating from a given target path of monetary growth. Most central banks have abandoned this policy, preferring to target the inflation rate itself.

Inflation targeting makes no sense in a model in which we still assume prices are fixed. We introduce inflation targeting in Chapter 21. In this chapter, we assume instead that the central bank pursues a **monetary target**. This is a good way to introduce many key ideas, and is useful in understanding how monetary policy was set in the 1980s before inflation targeting became popular.

Following a **monetary target**, the central banks adjust interest rates to maintain the quantity of money demanded in line with the target path for money supply.

We now combine our analysis of the goods market and money market to examine interest rates and output simultaneously. Chapters 16 and 17 analysed short-run equilibrium output using a diagram plotting income against demand. Since we now wish to keep track of interest rates explicitly, we need a new diagram.

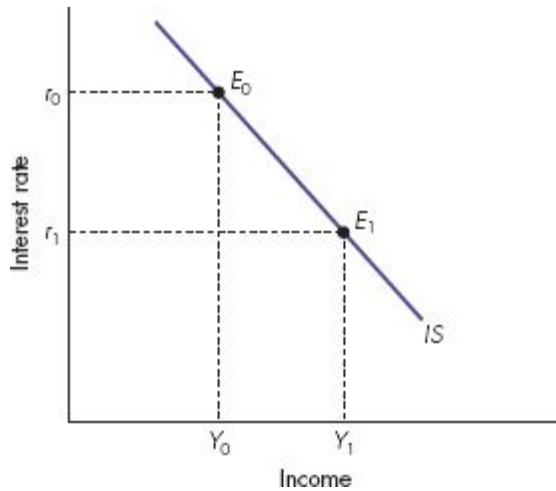
20.2 The *IS–LM* model

We consider *combinations* of income and interest rates that lead to equilibrium in each of the two markets – output and money – and thus determine the unique combination of income and interest rates yielding equilibrium in both markets at the same time.

The ***IS* schedule** shows combinations of income and interest rates at which aggregate demand equals actual output.

The *IS* schedule: goods market equilibrium

The goods market is in equilibrium when aggregate demand equals actual income. Hence, as shorthand, the combinations of interest rates and income compatible with short-run equilibrium in the goods market is called the *IS* schedule.¹



The *IS* schedule shows how a change in interest rates affects aggregate demand and short-run equilibrium output. A lower interest rate boosts demand and output. Anything else affecting aggregate demand shifts the *IS* schedule.

Figure 20.1 The *IS* schedule

Figure 20.1 shows the *IS* schedule. It is drawn for a given level of present and future government spending, a given level of present and future taxes and given present beliefs about future output and income. Holding these constant, lower interest rates increase both investment and consumption demand. At an interest rate r_1 , aggregate demand and short-run equilibrium output Y_1 are higher than their level Y_0 when the interest rate is r_0 .

Changes in interest rates move the goods market along a given *IS* curve. Anything else that affects aggregate demand is shown as a shift in the *IS* schedule.

The slope of the *IS* schedule

The *IS* schedule slopes down. Lower interest rates boost aggregate demand and output. The *slope* of the *IS* schedule reflects the sensitivity of aggregate demand to interest rates. If demand is very sensitive to

interest rates, the *IS* schedule is flat. A small change in interest rates causes a large change in aggregate demand for output. Conversely, if output demand is insensitive to interest rates, the *IS* schedule is steep. Changes in interest rates have only a small effect on aggregate demand for output.

Although, for simplicity, we often show the *IS* schedule as a *straight* line, it may in fact get steeper as we move downwards along it – there are probably diminishing returns to the effect of lower interest rates in boosting aggregate demand. We therefore sometimes refer to it as the *IS* curve, and never refer to it as the *IS* line.

Shifts in the *IS* schedule

Movements along the *IS* schedule show how interest rates affect aggregate demand and equilibrium output. Other changes in aggregate demand shift the *IS* schedule. For a *given* interest rate, more optimism about future profits raises investment demand. Higher expected future incomes raise consumption demand. Higher government spending adds directly to aggregate demand. Any of these, by raising aggregate demand at a given interest rate, raises equilibrium output at any interest rate – an *upward shift* in the *IS* schedule.

Conversely, greater pessimism among firms or households, a cut in government spending or higher tax rates cause a downward shift in the *IS* schedule.

CASE 20.1

FISCALLY CHALLENGED EUROZONE COUNTRIES

Since 2009 financial markets have been concerned about the fiscal solvency of a number of countries. Within the Eurozone , and setting aside the case of tiny Cyprus, attention has largely focused on Portugal, Italy, Greece and Spain – which had four characteristics: high government debt, high budget deficits, lack of international competitiveness and membership of the Eurozone. Their fiscal problems became the first real crisis of the Eurozone.

Despite sharing a common currency, interest yields on their bonds were many percentage points higher than interest yields on German

bonds, and their subsequent bond issues were expensive for their governments and taxpayers.

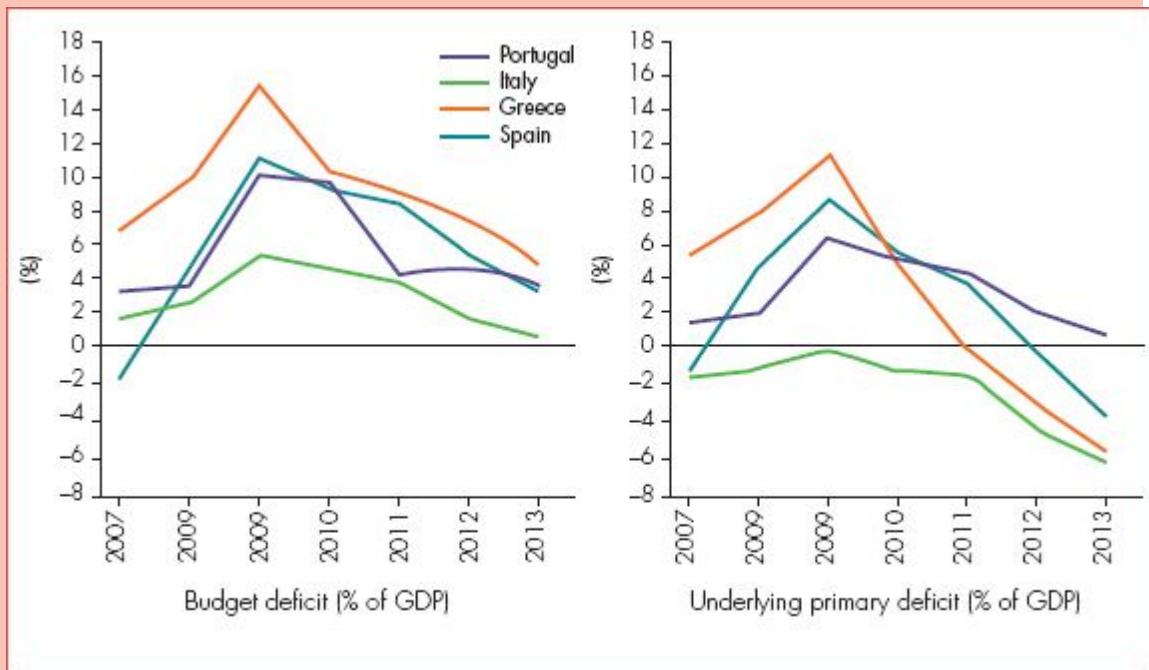
Optimists, such as Nobel Laureate Professor Joseph Stiglitz, argued that both interest rates and budget deficit indicators were misleading. If the crisis could be solved, risk premia embedded in their interest rates would evaporate as quickly as they had arisen. Budget deficits were also misleading because, as we saw in Chapter 17, the size of the budget deficit fluctuates with the level of output – in a slump tax revenue falls but, as output recovery occurs, tax revenue automatically rises again. Focusing on budget deficit data at the bottom of the slump gives a misleading impression of how bad the fiscal situation has become. Stiglitz therefore argued that Germany and France could help fiscally challenged members at little risk to themselves.

If the speculators could be defeated, the situation would correct itself without a need for default. In 2012, ECB President Mario Draghi undertook to purchase debts of these countries, justifying it by the need to fix the *transmission mechanism* of monetary policy within the Eurozone, whereby low ECB interest rates were supposed to feed through to low interest rates throughout the Eurozone. Many saw the move as a direct attempt to help the solvency of fiscally challenged countries, even though this was outside the remit of the ECB.

A fall in long-term interest rates has two effects on the *IS* schedule in such countries. First, it would move such countries down a given *IS* schedule, the direct effect of lower interest rates relevant for consumption and investment demand. Second, by easing the government's budget problems, it allows the government to have higher spending and/or lower tax rates than it would have had if it had continued to face higher borrowing costs. This fiscal expansion (relative to what might have been) leads to an *IS* schedule that is higher than it would have been in the presence of higher interest rates.

How large could such an effect be? The *primary* budget deficit shows the budget deficit *excluding* the interest payments on the debt. We can think of this as (a) an exact indicator of what would happen to the budget deficit if a country defaulted completely on its debt interest payments or (b) as a rough indicator of what would

happen to its deficit if interest rates fell from high penal levels to much lower levels. The Organization for Economic Cooperation and Development (OECD) makes estimates of the 'underlying primary deficit', not merely excluding interest payments but also using the hypothetical tax revenue that would accrue if output had been at 'normal' rather than 'actual' levels.



Source: OECD, *Economic Outlook*, November 2012.

The underlying primary deficit thus shows the state of the budget deficit once we strip out both debt problems and cyclical recession. It indicates how far off track the budget is in the long run. The charts show both the actual budget deficits of Portugal, Italy, Greece and Spain, and their underlying primary deficits.

The left-hand charts show all four countries gradually bringing their budget deficits under control after 2009, achieved by very austere fiscal policies that led to downward shifts in their *IS* curves, causing falls in output and higher unemployment. The right-hand chart shows the considerable progress that has been made with the underlying primary budgets, which are now in surplus or close to being in surplus. Inherited debts, penal interest rates and ongoing recessions are the cause of continuing problems. If these could be addressed somehow, fiscal policies themselves would not then be irresponsible or unsustainable.

Clearly, debtors should generally expect, and be expected, to repay their loans with interest. Sometimes, however, debtors are victims of circumstances not of their own making. Greece apart, none of the others had unsustainable policies when the crisis first occurred, and irresponsible behaviour in the US was the proximate cause of the crisis (albeit that property booms and irresponsible lending also occurred in most of Europe).

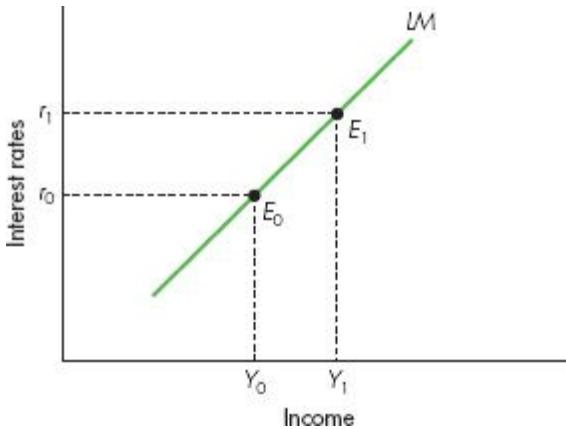
As the IMF and Eurozone ponder how much assistance now to offer countries in trouble, they need to consider two things. First, how much of the current predicament is a self-fulfilling crisis that could be resolved by long-term loans to allow struggling countries to enjoy lower interest rates, consequently loosen fiscal policy a little and shift *IS* schedules upwards? Second, how bad a precedent would be created by bailing out debtors in trouble, thereby undermining future incentives for prudent government behaviour?

The *LM* schedule: money market equilibrium

Pursuing a monetary target, the central bank endeavours to fix the money supply itself. In Figure 20.2, along the ***LM* schedule**, the demand for money (or liquidity L) equals the given supply of money (M , hence the shorthand LM).

The ***LM* schedule** shows combinations of interest rates and income that lead to money market equilibrium when the central bank pursues a given target for the nominal money supply.

The quantity of money demanded rises with level of output Y but falls with the level of the interest rate r . In money market equilibrium, money demand equals the given money supply. Hence if output rises from Y_0 to Y_1 – tending to raise the quantity of money demanded – money market equilibrium is restored only if interest rates rise from r_0 to r_1 , thereby reducing money demand back to the level of the given money supply. Figure 20.2 shows the upward-sloping schedule LM describing money market equilibrium. Higher output and income are accompanied by higher interest rates.



The *LM* schedule depicts money market equilibrium and is drawn for a given money supply. Higher income raises the quantity of money demanded. Only if interest rates are higher can the quantity of money demanded continue to equal the unchanged money supply.

Figure 20.2 The ***LM*** schedule

The slope of the schedule

The *LM* schedule slopes up. Higher output induces a higher interest rate to keep money demand in line with money supply. The more sensitive is money demand to income and output, the more the interest rate must change to maintain money market equilibrium, and the steeper is the *LM* schedule. Similarly, if money demand is not responsive to interest rates, it takes a big change in interest rates to offset output effects on money demand, and the *LM* schedule is steep. Conversely, the more money demand responds to interest rates and the less it responds to income, the flatter is the *LM* schedule.

As with the *IS* schedule, the *LM* schedule is almost certainly *not* a straight line. In particular, interests cannot fall below zero, so the *LM* schedule can never cross the horizontal axis. Again, we sometimes call it the *LM* curve, but never the *LM* line.

Shifts in the *LM* schedule

Movements along the schedule indicate interest rate changes to implement the *existing* policy as output changes. Shifts in the schedule reflect a *change* in monetary policy.

We draw an *LM* schedule for a *given* nominal money target. A rise in the target money supply means that money demand must also be increased to maintain money market equilibrium. This implies a rightward *shift* in the

LM schedule. Output is higher, or interest rates lower, raising money demand in line with the rise in real money supply.

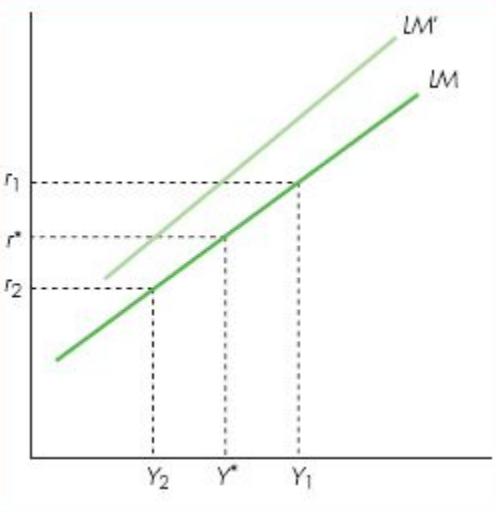
Conversely, a lower monetary target shifts the *LM* schedule to the left. Since money demand must also be reduced to preserve money market equilibrium, a higher interest rate is required at each income level. To sum up, moving along the *LM* schedule, higher interest rates need higher income to keep real money demand equal to the fixed supply. A higher (lower) target for money supply shifts the *LM* schedule to the right (left).

CONCEPT 20.1

A MODERN INTERPRETATION OF THE *LM* SCHEDULE

The *LM* schedule shows the relationship between interest rates and output implied by the monetary policy in force. Such a policy might be the pursuit of a fixed target for the quantity of money supplied.

We could instead interpret the *LM* schedule as a monetary policy in which the central bank deliberately sets higher interest rates when output is higher. This is consistent with a desire to stabilize output around its full capacity level. The steeper the *LM* schedule, the more aggressively the central bank 'leans into the wind' in order to offset deviations of output from full capacity. When aggregate demand falls and output is below potential output Y^* , the central bank reduces interest rates to r_2 , helping to mitigate the fall in aggregate demand and ensure output falls only to Y_2 . Conversely, when aggregate demand is higher than potential output, the central bank raises interest rates to r_1 , thereby restricting the increase in demand and output to Y_1 .



What is happening to the quantity of money supplied? The central bank is passively supplying whatever money is necessary to achieve money market equilibrium at the interest rate that the central bank wishes, given the level of output that is being produced. If the resulting money demand is large, the central bank simply ensures that the money supply is larger.

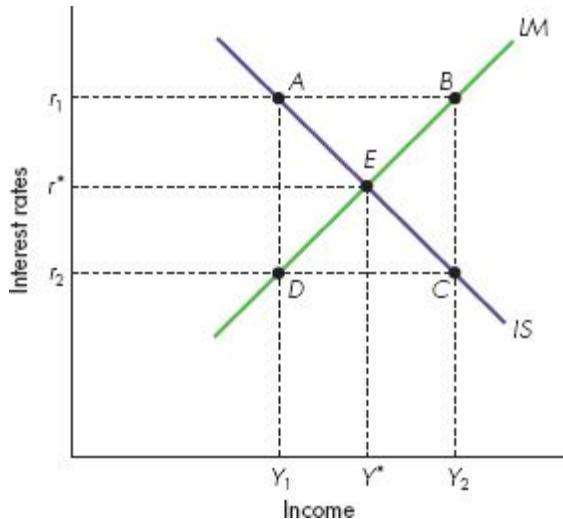
A steeper schedule means that interest rates are adjusted more, thereby achieving even greater stabilization of output, a flatter *LM* schedule means interest rates are adjusted less to stabilize output. A vertical *LM* schedule would stabilize output completely.

As well as the *slope* of the *LM* schedule, we also have to think about *shifts* in the schedule. An upward shift in the *LM* schedule, from *LM* to *LM'*, reflects a tighter monetary policy. At any output level, interest rates are higher under the new policy than under the previous one. For example, at output Y^* , the interest rate is now r_1 instead of r^* . The new monetary policy is more restrictive than the old one.

20.3 The *IS–LM* model in action

Figure 20.3 shows both the *IS* schedule, depicting combinations of income and interest rates consistent with goods market equilibrium, and the *LM* schedule, depicting combinations of interest rates and income consistent with money market equilibrium when the central bank's monetary policy rule is to pursue a fixed money supply target.

Equilibrium in both the money market and the output market is at point E , with an interest rate r^* and income level Y^* .



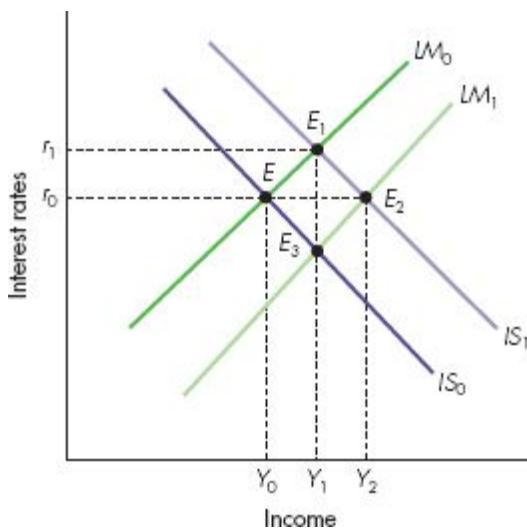
The goods market is in equilibrium at all points on the IS schedule. The money market is in equilibrium at all points on the LM schedule. Hence only at point E are both markets in equilibrium.

Figure 20.3 Equilibrium in the goods and money markets

Fiscal policy: shifting the IS schedule

Figure 20.4 shows the effect of a fiscal expansion that shifts the IS schedule from IS_0 to IS_1 . If unchanged monetary policy is shown by LM_0 , equilibrium moves from E to E_1 . Fiscal expansion leads to higher income but also higher interest rates. Higher output tends to increase the quantity of money demanded. Only higher interest rates prevent this from happening.

Fiscal contraction has the opposite effect. The IS schedule shifts to the left and output falls, tending to reduce money demand. Only lower interest rates restore money demand to the unchanged level of money supply, preserving money market equilibrium. In Figure 20.4, this is a move from E_1 to E when the IS schedule shifts down from IS_1 to IS_0 .



A fiscal expansion shifts the IS schedule from IS_0 to IS_1 but leaves the LM schedule unaltered at LM_0 . Equilibrium moves from E to E_1 . Output rises only from Y_0 to Y_1 because the output expansion induces a rise in interest rates from r_0 to r_1 that dampens the rise in aggregate demand. By accompanying the fiscal expansion with a monetary expansion from LM_0 to LM_1 , policy could make output rise to Y_2 . Fiscal expansion makes output rise more if monetary policy is loosened to keep interest rates unaltered.

Figure 20.4 Fiscal expansion shifts the IS schedule

Figure 20.4 makes three other points. First, **crowding out in the Keynesian model** is complete – extra government spending G leads to an equivalent reduction in consumption and investment ($C + I$), leaving output unaltered – only if the LM schedule is vertical. Then, an upward shift in the IS schedule raises interest rates but not income.

A fiscal stimulus to aggregate demand **crowds out in the Keynesian model** some private spending. Higher output induces higher interest rates that dampen the expansionary effect on aggregate demand.

In practice, the LM schedule is never completely vertical, which would occur only if it took an *infinite* rise in interest rates to offset the effect of slightly higher output on money demand. Since the LM schedule normally has a positive slope, fiscal expansion raises demand and output despite some induced rise in interest rates.

Second, fiscal policy is not the only possible autonomous change in aggregate demand. An increase in export demand would also shift the IS

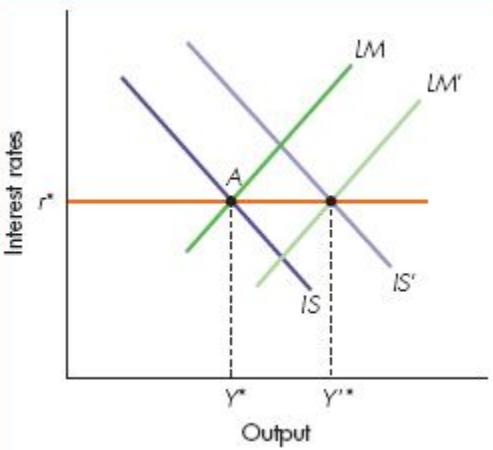
schedule to the right, again inducing higher output and higher interest rates. Movements *along* the *IS* schedule show the effect of interest rates. All other shifts in aggregate demand imply *shifts* in the *IS* schedule.

Third, Figure 20.4 shows what happens if fiscal expansion is *accompanied* by a looser monetary policy. Fiscal expansion shifts *IS* to the right, but monetary expansion – a higher money supply target – shifts *LM* to the right. It is possible to loosen monetary policy just enough to keep interest rates at their original level when income expands. Fiscal expansion then leads to a new equilibrium at E_2 , with interest rates unchanged at r_0 . Hence, the output effect of a fiscal expansion depends on the monetary policy in force. The more that monetary policy prevents a rise in interest rates, the more the fiscal expansion will lead to higher output.

CONCEPT 20.2

A HORIZONTAL *LM* SCHEDULE

If monetary policy is always adjusted to keep interest rates constant, we may as well view the *LM* schedule as horizontal at the target interest rate, as shown in the figure below. Suppose a fiscal expansion shifts the *IS* schedule from *IS* to *IS*₉. If the central bank loosens monetary policy from *LM* to *LM*₉, it can maintain interest rates at the original level despite the increase in output and money demand – it has simply increased money supply to match. The money supply is passively adjusted to whatever level of money is demanded at that interest rate. Shifts in the *IS* schedule no longer lead to crowding out because the money supply is adjusted to prevent interest rates from changing. Instead of depicting monetary policy as a whole potential set of parallel *LM* schedules, it is easier just to summarize it by the horizontal line at height r^* . Whatever happens to the *IS* schedule, monetary policy will then be adjusted to maintain the interest rate at a constant level.



In Chapter 24 we show that defending a fixed exchange rate may require a constant interest rate and hence a **horizontal LM schedule**. Hence, small countries within the Eurozone may face a horizontal LM schedule. The European Central Bank (ECB) sets an interest rate for the whole Eurozone, and countries such as Belgium and the Netherlands have to take this interest rate as given. Germany is a little luckier – as the largest country within the Eurozone, German economic conditions tend to affect the ECB's decisions regarding the interest rate within it. Most of the time, euro interest rates are higher when German output is higher; Germany enjoys an LM curve that slopes upwards. Upward-sloping LM curves mean that monetary policy is acting to stabilize output fluctuations by raising interest rates when that country's output is higher.

A **horizontal LM schedule** implies the money supply is adjusted to keep interest rates constant.

Monetary expansion: shifting the LM schedule

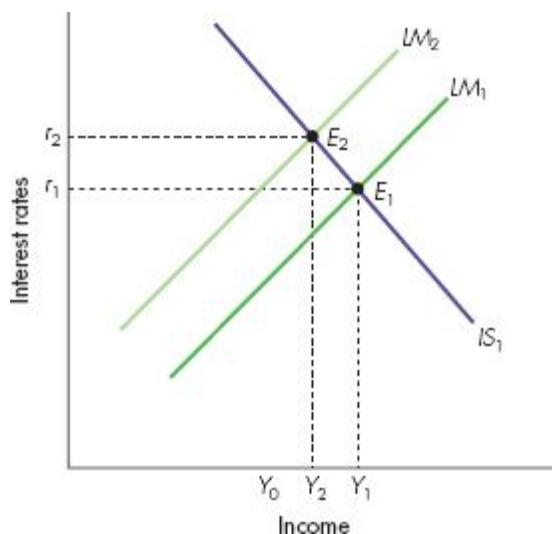
Similarly, beginning from E in Figure 20.4, an increase in the target money supply shifts the LM schedule from LM_0 to LM_1 : for any income, it requires lower interest rates to help raise money demand in line with the new higher money supply. Lower interest rates also boost income, which also helps raise money demand. Equilibrium moves from E to E_3 .

Conversely, a reduction in the target money supply shifts the LM schedule to the left, leading to higher interest rates but lower output.

20.4 Shocks to money demand

In the last three decades, competition between banks has increased dramatically, raising interest rates paid on deposits. Since the opportunity cost of holding money in a bank deposit is only the differential between the deposit interest rate and the higher interest rate available on other financial assets, changes in banking competition change the opportunity cost of holding money *at any market interest rate* r . Conversely, since the financial crash, banks have been desperate to increase profit margins in order to rebuild capital reserves. The spread between deposit interest rates and market interest rates has widened sharply.

We draw an LM schedule for a given nominal money target. Greater banking competition raises money demand at every combination of output and interest rates. To keep money demand in line with the unchanged supply, either output must fall or interest rates must rise. The LM schedule *shifts* left. Conversely, if spreads widen and the opportunity cost of holding money increases, the LM schedule shifts to the right.



An undetected shift in money demand shifts the LM curve, leading to a different equilibrium from that which the central bank intended when deciding what level of the money supply to target.

Figure 20.5 An unexpected rise in money demand

Figure 20.4 showed how changes in money *supply* shift the LM schedule under monetary targeting. We have now discovered that changes in money *demand*, other than those caused by changes in output and interest rates, also shift the LM schedule under monetary targeting.

In Figure 20.5, LM_1 corresponds to 'low' money demand and LM_2 to 'high' money demand. Suppose money demand increases but the central bank is not yet aware of the change. In choosing what monetary target to set, the central bank is expecting the schedule LM_1 , which will place the economy at E_1 . In fact, because of the undetected shift in money demand, the actual out-turn is at E_2 , not at all what monetary policy intended when it decided what monetary target to set.

In practice, this helps explain why monetary targets were gradually abandoned by many central banks. When money demand was predictable, monetary targets worked fine. As the financial sector became more sophisticated, more competitive and more volatile, monetary targets were gradually abandoned as the basis for the monetary policy rule.

Moreover, as we saw in Chapter 18, the bank deposit multiplier can be highly unstable. This means that a given quantity of narrow money can imply very different quantities of broad money. It is much simpler for central banks to decide what interest rate they wish to set, and then passively supply whatever narrow money is necessary to get whatever quantity of broad money is needed for money market equilibrium at that interest rate.

MATHS 20.1

THE MONETARY FISCAL MIX

Consider the model:

$$Y = A - br \quad \text{IS schedule} \quad A, b > 0 \quad (1a)$$

$$Y = D + er \quad \text{LM schedule} \quad e > 0, 0 < D < A \quad (1b)$$

Hence, $Y = A - br = D + er$, so in short-run equilibrium:

$$r = (A - D)/(b + e) \quad Y = (Ae + bD)/(b + e) \quad (2)$$

Thus, for example, a rise in autonomous aggregate demand will lead to an increase in short-run equilibrium output and interest rates.

To understand the LM schedule in more detail, we can use the money market equations:

$$\begin{array}{lll} M = fY - hr & \text{Money demand} & f, h > 0 \\ M = mR & \text{Deposit multiplier} & m > 0 \end{array}$$

Hence,

$$Y = [mR + hr]/f = [mR/f] + (h/f)r \quad (3)$$

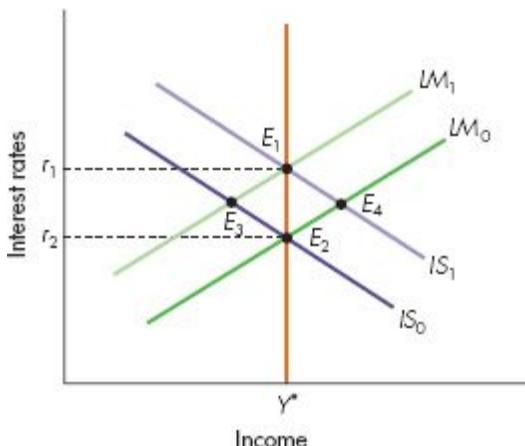
Comparing this with equation (1b), we can see that the constant D in the LM schedule, which determines how far to the right the LM schedule lies, is just $[mR/f]$ and will increase if the central bank supplies more reserves R , if banks raise the deposit multiplier m , or if money demand becomes more sensitive to income via the parameter f . Moreover, the slope of the schedule, which depends on e in equation (1b), simply depends on the parameters h and f .

20.5 The policy mix

Fiscal policy is government decisions about tax rates and spending levels. Changes in fiscal policy shift the IS schedule. Changes in monetary policy shift the LM schedule.

We now explore consequences of different IS and LM schedules (different monetary and fiscal policies). Budget deficits can be financed by printing money or by borrowing. In the latter case, there is no short-run connection between monetary and fiscal policy provided the government is solvent and can borrow any reasonable amount that it wishes. The government can then pursue independent monetary and fiscal policies.

Although both fiscal and monetary policy can alter aggregate demand, the two policies are not interchangeable. They affect aggregate demand through different routes and have different implications for the *composition* of aggregate demand.



The target income Y can be attained by easy fiscal policy and tight monetary policy. Equilibrium at E_1 , the intersection of LM_1 and IS_1 , implies high interest rates r_1 and a low share of private sector investment and consumption in GNP. Alternatively, with easy monetary policy and tight fiscal policy, equilibrium at E_2 , the intersection of LM_0 and IS_0 , still attains the target income but at lower interest rates r_2 . The share of private sector investment and consumption in GNP will be higher than at E_1 .

Figure 20.6 The policy mix

Figure 20.6 shows the mix of monetary and fiscal policy. There are two ways to stabilize income at Y^* . First, there is expansionary or *easy* fiscal policy (high government spending or low tax rates). This leads to a high IS schedule, IS_1 . To keep income in check with such an expansionary fiscal policy, *tight* monetary policy is needed. With a low money supply target, the schedule LM_1 is far to the left.

Equilibrium at E_1 achieves an output Y^* but also a high interest rate r_1 . With high government spending, private demand must be kept in check. The mix of easy fiscal policy and tight monetary policy implies government spending G is a big share of national income Y^* but private spending ($C + I$) a small share.

Alternatively, the government can adopt a tight fiscal policy (a low IS_0 schedule) and an easy monetary policy (LM_0 far to the right). The target income Y^* is now attained with a lower interest rate r_2 at the equilibrium E_2 . With easy monetary policy and tight fiscal policy, the share of private expenditure ($C + I$) is higher, and the share of government expenditure is

lower, than at E_1 . With lower interest rates, there is less crowding out of private expenditure.

Of course, easy monetary policy *and* easy fiscal policy together are highly expansionary. With the schedules IS_1 and LM_0 the equilibrium in Figure 20.6 is at E_4 . Income is well above Y^* . Conversely, with tight monetary policy and tight fiscal policy, and schedules LM_1 and IS_0 , equilibrium is at E_3 , with income well below Y^* .

What should determine the mix of fiscal and monetary policy? In the long run, the government may care not just about keeping output close to potential output, but also about raising potential output. High investment increases the capital stock more quickly, giving workers more equipment with which to work and raising their productivity. Governments interested in long-run growth may choose a tight fiscal policy and an easy monetary policy. Conversely, if governments are politically weak and unable to resist demands for high government spending to pay off various factions, fiscal policy will be loose and a tight monetary policy is needed to keep aggregate demand in line with potential output.

Solvency concerns also affect the feasible monetary–fiscal mix. When financial markets panic about the solvency of particular governments, further bond-financed fiscal expansion may not be possible. Nobody will buy the bonds. The more difficult it is to meet aims through one policy, the more desirable it is for the other policy to do the work. The problem for Eurozone countries such as Greece was that, as members of the Eurozone, they had surrendered the ability to use an independent monetary policy to target their needs alone.

One final point about the monetary–fiscal mix: changing fiscal policy takes time whereas monetary policy can be changed very quickly.

ACTIVITY 20.1

MONETARY OR FISCAL POLICY?

Concept 17.1 noted some reasons why fiscal policy may not be ideal for short-run management of aggregate demand. Some of these reasons – for example, problems in diagnosing where the economy is and forecasting where it might go if policy is left unchanged – apply just as much to monetary policy as to fiscal

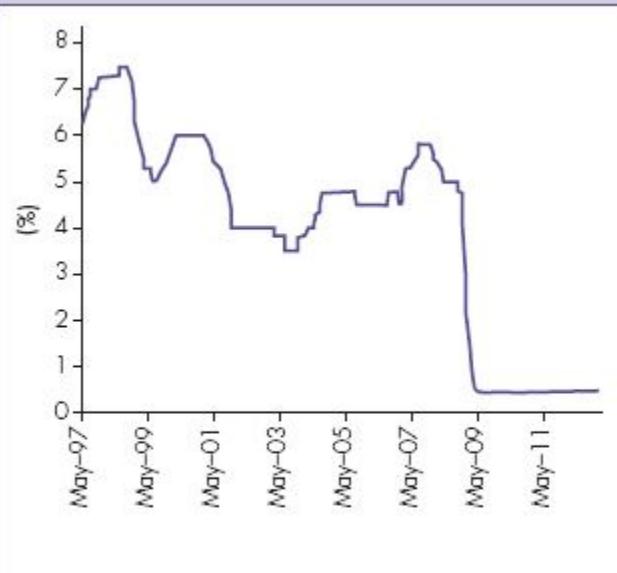
policy. However, two problems are often thought to make fiscal policy less suitable for short-run variation.

First, fiscal policy is difficult to change quickly. Rapid changes in hospital building or in tax rates are more costly than rapid changes in interest rates. Financial markets are accustomed to asset prices changing quickly. Second, it is politically easy to loosen fiscal policy but politically much more difficult to tighten it again later. For this reason, the most important source of short-term movements in fiscal policy is the operation of automatic stabilizers. Since tax rates are not changing, no visible decisions are being made to which voters could object. Yet tax revenue is varying with output.

It used also to be politically difficult to tighten monetary policy. For example, people (voters!) who have borrowed to buy a house get upset when interest rates rise sharply. The main reason that most countries have made their central banks independent of political control in decisions about interest rates is precisely to take the politics out of monetary policy. Nowadays, interest rates can and do change rapidly, in both directions, though usually by very small amounts. The chart shows UK data from 1997, when the Bank of England formally became independent, until 2012.

The figure confirms how aggressively interest rates were reduced in 2008/09 once the magnitude of the financial crisis became apparent.

What does this imply about the monetary–fiscal mix? The budget deficit was already high by 2009 because of the need to bail out banks. Further fiscal expansion threatened to create unsustainable levels of government debt that would have been difficult to repay. In these circumstances, monetary policy had to provide as much stimulus as possible, both because a large stimulus was needed and because fiscal policy was already overstretched.



UK bank rate 1997–2012

Source: www.bankofengland.co.uk.

By March 2009 the Bank of England's official rate had been cut to 0.5 per cent and could hardly go much lower. It remained there for the next three years. As we saw in Chapter 19, the Bank then embarked on a programme of quantitative easing, partly to offset the collapse of the bank deposit multiplier and partly in the hope of providing additional stimulus by bidding down the prices of long bonds, thereby reducing long-term interest rates.

Questions

- During which periods since 1997 was the Bank of England most worried about inflation? Why do you think this?
- Were interest rates changed in response to inflation at the time or to the prospect of inflation a year or two subsequently? Why do you think this?
- Should the Bank worry about changing its mind, raising interest rates but sometimes lowering them shortly afterwards if necessary, or should it act more slowly so that it rarely has to reverse its recent decisions?
- Suppose an output slump leads to a period of negative inflation. What is the lowest possible nominal interest rate? What would then happen to real interest rates?

To check your answers to these questions, go to page 680.

20.6 The effect of future taxes

Chapter 19 argued that consumption demand reflects both *current* disposable income and expected *future* disposable income. Two hundred years ago, the English economist David Ricardo noticed a striking implication. Suppose the path of government purchases G is fixed over time. What path of taxes over time finances this spending?

The government can lend and borrow. In some years, its spending may exceed taxes; in other years, taxes must then exceed spending.

For a given planned path of spending, and a suitable planned path of tax revenue, the government cuts taxes this year, and pays for it by borrowing – hence achieving **government solvency**. It sells bonds. The tax cut is a fiscal expansion that boosts aggregate demand. Right?

Government solvency requires that the present value of current and future tax revenue equals the present value of current and future spending plus any initial net debts.

If the tax cut is £1 billion, this is also the value of bonds issued to finance it. The market value of bonds is the present value of future income to bondholders. By assumption, the path of government spending is fixed. Hence, interest payments to bondholders must be financed by higher taxes in the future. £1 billion is the value of the tax cut, *and* the value of the new bonds, *and* the present value of the extra future taxes. The private sector gets a handout today (a tax cut) offset by a future penalty (higher taxes) of identical present value. The private sector is neither richer nor poorer. Its desired spending should not change. Today's tax cut has no effect on aggregate demand because it is matched by the *prospect* of higher future taxes.

Equivalently, the fall in government saving (larger deficit today) is exactly offset by a rise in private saving: private spending is unaltered, and larger disposable incomes (because of the tax cut) go entirely in extra saving (to pay for the future taxes).

Some people getting tax cuts today will die before future taxes arrive. But suppose these people have children and care about them. After a tax cut today, parents save more to bequeath extra money to their children, or

grandchildren, to pay the higher future taxes. The extra disposable income is saved to raise the bequest for future generations.

Ricardian equivalence does not deny that road-building, financed by higher taxes, affects aggregate demand. Government spending always has real effects. Rather, for a *given* path of real government spending, it may not matter *when* people pay for it. Ricardo himself thought the equivalence hypothesis would not hold in the real world. Economists are still arguing about the extent to which **Ricardian equivalence** should hold.

Ricardian equivalence says that it does not matter when a government finances a given long-run spending programme. Tax cuts today do not affect private spending if, in present value terms, future taxes have to rise to match.

Why Ricardian equivalence is too strong

There are three reasons why the tax cuts today *do* stimulate demand a bit even if future taxes are correspondingly higher. First, people without children get the benefit of tax cuts without paying the full burden of higher future taxes in the distant future. They spend more at once.

Second, by reducing marginal tax rates and distortions, tax cuts may increase potential output and raise income. Expecting higher incomes, people spend more immediately.

Third, solvent governments can borrow at a low interest rate. Ricardian equivalence holds only if we can all borrow as easily as the government. If only! Households and firms are individually riskier than governments. Private people have no residual power to tax or print money when things go wrong. Hence, lenders charge private borrowers a higher rate of interest, and may refuse to lend at all.

Now do the sums again. £1 billion is the value of the tax cut, the extra government bonds and the present value of extra tax payments *discounted at the interest rate faced by the government*. We face a higher interest rate when we try to borrow. *As viewed by us, the present value of our extra future taxes is less than £1 billion because we discount at a higher interest rate.*

The tax cut is a fiscal expansion because in effect the government borrows on the good terms it enjoys, then lends to us at better terms than the capital market. It gives us a loan, tax cuts today, which we repay later in higher taxes. But we are charged the government's low interest rate for our loan. We are better off and spend more. Aggregate demand increases.

Theory and evidence suggest that complete Ricardian equivalence is too extreme to fit the real world. Tax cuts do boost aggregate demand today (though higher future taxes will reduce demand at some future date). Ricardian equivalence is not completely right, but not completely wrong. Expectations of future conditions affect current behaviour. Private saving rises a bit when public saving falls. The private sector does substitute between present and future, despite obstacles to doing this easily. These obstacles make consumption demand more sensitive to current disposable income than it would be if borrowing were easy and only permanent income mattered.

Current demand by firms and households depends both on current fiscal policy and expected future fiscal policy. Since one does not fully offset the other, for simplicity we can look at current fiscal policy in isolation. We need to remember only that some of its quantitative effects will be smaller if people expect fiscal policy to have to be reversed at some future date.

If Ricardian equivalence held exactly, government efforts to prop up aggregate demand in the aftermath of a financial crisis by running budget deficits would have been a waste of time. But understanding the trade-off between the present and the future allows three insights into the events since 2008:

1. It was precisely in 2009 that banks were most scared to lend and the private sector had so much difficulty borrowing. People who cannot borrow at all have to use an infinite discount rate in discounting the future. These are the circumstances in which bond-financed tax cuts are most powerful. They increase private sector liquidity at the critical time. The government can borrow on better terms than its individual citizens can.
2. Conversely, as governments get closer to the limits of what they can easily borrow and guarantee to repay, the differential between private sector and government creditworthiness narrows. At some future point, deficit-financed tax cuts (or other subsidies to the private sector) lose most of their power.

3. Two specific measures adopted in 2009 – a temporary VAT cut and a temporary subsidy to scrappage of old cars –worked not by increasing the permanent income of households but by persuading them to bring forward spending from the future to the present. This is great when the measures are first introduced (and helped explain positive UK output growth by the fourth quarter of 2009), but then leads to a corresponding fall in demand when we get to the future from which the spending has been brought forward. All that has occurred is a retiming of spending, not an increase in the entire path of spending.

Governments wish to bring spending forward because of the effect on confidence. If this can be established, by ending the downward spiral, growth may become strong enough to cope with the future lack of spending for a while. Given that most European economies were back in recession by 2012, the temporary restoration of confidence did not then set their economies on sustainable growth paths. Once growth became fragile, confidence soon evaporated again.

20.7 Demand management revisited

In the last five chapters we have studied how aggregate demand determines output and employment. Fiscal and monetary policy can manage aggregate demand, aiming to keep the economy close to its full-employment level. In periods of recession, when aggregate demand is insufficient, monetary and fiscal expansion can boost demand, output and employment.

Demand management uses monetary and fiscal policy to stabilize output near potential output.

Thus far, we have treated the price level as given. If the price level can change, boosting demand may lead not to higher output but to higher prices. In the next chapter, we begin the study of prices and inflation. In so doing, we introduce aggregate supply, and hence the balance between aggregate supply and aggregate demand.

However, you have now completed the first stage of macroeconomics, learning how to analyse the demand side of the economy. Even after mastering the analysis of supply, adjustment and price behaviour, the

demand analysis of the last few chapters remains a key part of the story, especially in the short run.

Summary

- A **given fiscal policy** means a given path of government spending and tax rates. A **given monetary policy** must specify the implicit **monetary policy rule** by which interest rates are set. In this chapter, we assume that is to achieve a given **money supply target**
- The ***IS schedule*** shows combinations of interest rates and output compatible with short-run equilibrium output in the goods market. Lower interest rates boost demand and output. Other causes of shifts in demand are shown as shifts in the *IS* schedule.
- The ***LM schedule*** shows combinations of interest rates and output compatible with money market equilibrium when the central bank pursues a money supply target. Higher output is associated with higher interest rates to maintain the equality of money supply and money demand.
- The intersection of *IS* and *LM* schedules shows simultaneous equilibrium in both goods and money markets, jointly determining output and interest rates.
- With a given monetary policy, a fiscal expansion increases output, money demand and interest rates, thus **crowding out** or partially displacing private consumption and investment demand.
- For a given fiscal policy, a **monetary expansion** leads to lower interest rates and higher output.
- The **mix of monetary and fiscal policy** affects the equilibrium interest rate as well as the level of output.
- **Ricardian equivalence** says that, for a given present value of government spending, the private sector does not care *when* this is financed by taxes, since the total present value of taxes is the same. A

tax cut today has no effect on aggregate demand since people anticipate higher future taxes to finance the extra debt interest.

- Ricardian equivalence is true only under extreme assumptions not generally true in practice. Tax cuts today do have some effect today. This effect is dampened by the knowledge that, unless government spending is also cut, future taxes will have to rise.
- **Demand management** helps stabilize output. Fiscal policy may be difficult to adjust quickly, and may be difficult politically to reverse later: much of its impact on aggregate demand thus arises through **automatic stabilizers** with an unchanged fiscal policy.

Review questions



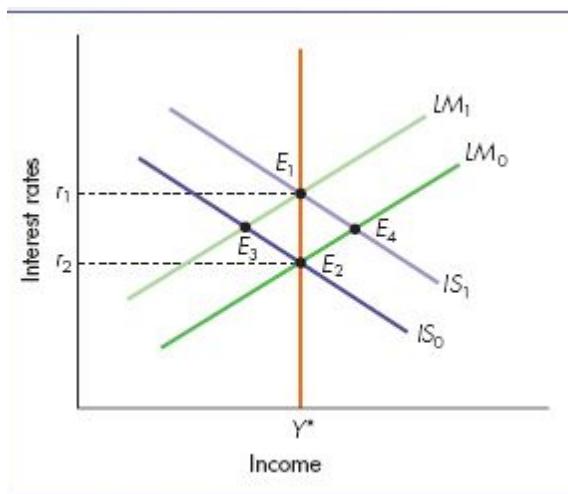
EASY

- 1 A small country that has adopted the euro must accept the single interest rate set for the whole of Euroland. Draw the *LM* schedule relating the interest rate to that country's national output. Why would this schedule ever shift?
- 2 Which of the following is correct? Other things equal, high output and high interest rates imply: (a) loose monetary policy and loose fiscal policy, (b) tight fiscal policy and tight monetary policy, (c) tight fiscal policy and loose monetary policy, (d) none of the above.
- 3 Why do people usually save a 'one-off income tax rebate'?
- 4 Suppose a government lived for ever and never broke its promises. Facing a large budget deficit and large government debt today, should the bond market be confident that any level of initial debt will be repaid provided the government pledges to do so sometime in the future?
- 5 **Common fallacies** Why are the following statements wrong? (a) If tax rates never change, fiscal policy cannot stabilize output. (b) Higher government spending makes interest rates rise, which could cut aggregate demand by more than the rise in government spending. (c) Future policy cannot affect present behaviour.

MEDIUM

- 6 Suppose mortgage lenders issued 20-year loans at fixed interest rates.

- a. How would short-term changes in interest rates impact households with a mortgage?
- b. Would the Bank of England have to change interest rates by more or by less to have the same effect on aggregate demand at present?
- 7 Suppose the European Central Bank has a monetary policy rule that relates Euroland's interest rate to total output in Euroland. If a small Euroland country's output is perfectly correlated with the output of all Euroland, draw the *LM* schedule (a) for Euroland and (b) for the small member country.
- 8 In 2010, having accumulated substantial government debt owed by foreigners and having a large budget deficit, Greece lost the confidence of its international creditors and was given a loan on condition that it embarked on substantial fiscal tightening.
- a. Illustrate these changes, using the *IS–LM* diagram, recognizing that Greece is a Eurozone member. (b) Suppose Greece left the Eurozone; what, if anything, would be different?
- 9 For each of these shocks, say whether it shifts the *IS* schedule or the *LM* schedule, and in which direction: (a) an expected future fiscal expansion, (b) a higher money supply target, (c) a rise in money demand causing higher interest rates being paid by banks on bank deposits.
- 10 Suppose monetary policy raises nominal interest rates by 0.8 every time inflation rises by 1. (a) How do you expect the central bank gets on in stabilizing inflation around a low level? (b) Suppose inflation is nevertheless low and stable: how might you explain this outcome?
- 11 The diagram below shows the working of the *IS–LM* model where a mix of monetary and fiscal policy is used to stabilize income at Y^* .



If the government adopts a tight monetary policy and an easy fiscal policy, how can income be stabilized at Y^* ? How can income be stabilized with an easy monetary policy and a tight fiscal policy? Can income be stabilized at Y^* if the government adopted a tight monetary policy and tight fiscal policy?

HARD

- |2 **Essay question** ‘If households can lend and borrow easily, their consumption and saving decisions simply offset anticipated future tax changes. The principal power of taxation policy to influence aggregate demand arises because households in practice face difficulties borrowing what would be required to implement Ricardian equivalence.’ Discuss.
- |3 Use the *IS–LM* diagram to depict (a) the start of a financial crash in which confidence evaporates in the private sector and the banking system and (b) a subsequent policy of quantitative easing.
- |4 Imagine a world of only two periods and zero interest rates. A consumer’s income is 100 in each period, taxes are 50 each period, permanent disposable income is therefore 50, and consumption is 50 per period (since the world ends after period 2). The government now offers a tax cut of 10 in period 1, financed by government borrowing that will be repaid in period 2. (a) Since interest rates are zero, by how much must the government raise taxes in period 2 in order to pay off its loan in full? (b) What is the consumer’s disposable income now in each period? (c) Since the interest rate is zero, what is permanent disposable income? (d) What is the effect on consumption decisions? (e) If the government pays zero interest on loans, but the consumer pays 10 per cent interest, how is permanent income affected? (f) What is now the effect of the temporary tax cut?

1 The name *IS* schedule derives from the fact that, in the simplest model without either a government or a foreign sector, equilibrium income is where planned investment I equals planned saving S . However, the *IS* schedule – the combinations of income and interest rates consistent with equilibrium income – can be constructed for models including the government and foreign sector as well.

CHAPTER 21

Aggregate supply, prices and adjustment to shocks

Learning outcomes

By the end of this chapter, you should be able to:

- 1 understand inflation targets for monetary policy
- 2 recognise the ii schedule
- 3 describe how inflation affects aggregate demand
- 4 understand aggregate supply in the classical model
- 5 analyse the equilibrium inflation rate
- 6 understand complete crowding out in the classical model
- 7 recognise why wage adjustment may be slow
- 8 analyse short-run aggregate supply
- 9 understand temporary and permanent supply shocks
- 10 describe how monetary policy reacts to demand and supply shocks
- 11 recognise flexible inflation targets
- 12 identify a Taylor rule

Keynesian models suggest that higher aggregate demand always raises output. However, with finite resources, the economy cannot expand output indefinitely. We now introduce aggregate supply – firms' willingness and ability to produce – and show how demand and supply together determine output. Aggregate demand reflects the interaction of the markets for goods and money. Aggregate supply reflects the interaction of the markets for goods and labour.

Introducing supply means that we abandon the simplifying assumption that output is determined by demand alone. With both supply and demand, we can also explain what determines prices. We no longer need to assume that prices are given. And since inflation is simply the growth of prices from period to period, a model of prices is also a model of inflation. This allows us to represent monetary policy as inflation targeting, the policy rule actually followed by most central banks today.

The **classical model** of macroeconomics assumes wages and prices are completely flexible.

To get started, we swap the Keynesian extreme, with fixed wages and prices, for the opposite extreme, full wage and price flexibility. In the **classical model**, the economy is always at full capacity. Any deviation of output from full capacity causes instant price and wage changes to restore output to potential output. In the classical model, monetary and fiscal policies affect prices but not output.

In the short run, until prices and wages adjust, the Keynesian model is relevant. In the long run, once all prices and wages have adjusted, the classical model is relevant. We study the transition from the Keynesian short run to the classical long run.

21.1 Inflation and aggregate demand

Inflation is the growth rate of the price level of aggregate output.

If a central bank behaves predictably, its behaviour can be modelled. Many central banks used to set interest rates to achieve a desired path for nominal money. Knowing the money demand schedule – the relation between interest rates, income and desired money holdings – the central bank forecast income, and chose interest rates to ensure that the quantity of money demanded equalled that it wished to supply.

Money demand became increasingly volatile and unpredictable because of financial innovation and changing competition between banks. Both affected the interest rates paid on bank deposits, and hence the opportunity cost of holding money. If central banks could not predict money demand, they did not know at what level to set interest rates in order to achieve desired money supply.

With an **inflation target**, the central bank adjusts interest rates to try to keep inflation close to the target inflation rate.

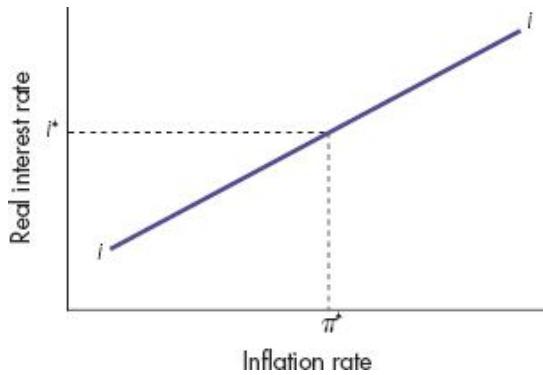
Monetary targeting was first abandoned in the countries experiencing the greatest financial innovation and banking competition: the US and the UK. Germany, initially more cautious about financial deregulation, retained monetary targeting for longer. Today, most central banks pursue an **inflation target**. Because of its strong German heritage, the European Central Bank pays attention both to inflation and monetary targets, a slightly uneasy compromise.

Target annual inflation π^* varies from country to country, but is usually around 2 per cent. Why not target zero inflation? Policy makers are keen to avoid *deflation*

(negative inflation). Even if the nominal interest rate r is reduced to zero, the real interest rate i , which is simply $(r - \pi)$, can be large if inflation π is large but negative. High real interest rates cause further contraction, make inflation more negative still, thus making real interest rates even higher. If nominal interest rates have already been reduced to zero, monetary policy can do little to combat shrinking aggregate demand.

To avoid this black hole, a positive inflation target leaves a margin of error. If an unforeseen shock reduces inflation, there is still time for the central bank to boost the economy before it gets too close to a deflationary spiral.

Figure 21.1 shows how monetary policy works when interest rates are set to achieve an inflation target. If inflation is high, the central bank ensures that real interest rates are high, reducing aggregate demand, and putting downward pressure on inflation.



When inflation is above (below) the target π^* , real interest rates are set higher (lower) than normal.

Along the schedule ii , a given monetary policy is being pursued. If the inflation target is π^* , the corresponding real interest rate will be i^* .

Figure 21.1 Interest rates and inflation targeting

Under inflation targeting, the *ii schedule* shows that, at higher inflation rates, the central bank will wish to have higher real interest rates.

With a vertical *ii schedule*, inflation would be completely stabilized at its target rate π . If inflation started to rise, real interest rates would be raised as much as necessary to restore inflation to its target level. Conversely, if inflation started to fall, real interest rates would be reduced to the level needed to restore the inflation to target.

Such a monetary policy would be too aggressive. Later in the chapter we show why some of its side effects would be undesirable. The *ii* schedule in Figure 21.1 shows more moderate intervention. When inflation is too high, the central bank

raises real interest rates a bit; when inflation is too low, real interest rates are reduced a bit.

The central bank sets the **nominal interest rate r** not the **real interest rate i** .

Although the central bank is interested in the real interest rate, which affects aggregate demand, the central bank does not directly control the price of output or the inflation rate. Hence, to achieve the ii schedule of Figure 21.1, the central bank first forecasts inflation, then sets a **nominal interest rate r** to achieve the desired **real interest rate i** (which is just $r - \pi$).

The central bank sets the **nominal interest rate r** not the **real interest rate i** .

The **aggregate demand schedule AD** shows how inflation affects aggregate demand when the interest rate is set in pursuit of an inflation target.

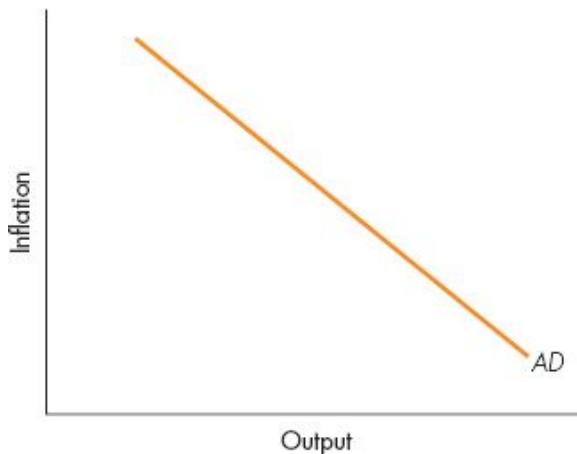
An implication of Figure 21.1 is that higher inflation leads to a *larger* rise in the nominal interest rate, so that the real interest rate is higher when inflation is higher.¹

A particular ii schedule is a particular monetary policy. Moving along the schedule, the central bank adjusts interest rates to inflation according to the policy rule already adopted. The policy is not changing. *Changes* in monetary policy are shown by *shifts* in the schedule. A looser monetary policy means a downward shift in the ii schedule; that is, a lower interest rate at each possible inflation rate. A tighter monetary policy shifts the ii schedule upwards; that is, a higher interest rate at each possible inflation rate.

If π^* is the inflation target, the chosen height of the ii schedule determines the corresponding real interest rate i^* when the inflation target is being met. A tighter monetary policy (higher ii schedule) means accepting a higher real interest rate i^* at the given inflation target π^* , or a lower inflation target at the same real interest rate i^{*0} .

Figure 21.2 shows the level of aggregate demand for output when interest rates obey the ii schedule implied by inflation targeting. Movements *along* the **aggregate demand schedule AD** show how inflation makes the central bank alter real interest rates and thus aggregate demand.² The AD schedule is flat when (a) interest rate decisions react a lot to inflation and (b) interest rates have a big effect on aggregate demand. The AD is steep when (a) interest rate decisions do not respond much to inflation and (b) changes in interest rates have a small effect on aggregate demand.

Shifts in AD reflect all other shifts in aggregate demand *not* caused by the effect of inflation on interest rate decisions. Thus, AD shifts up if fiscal policy eases, net exports rise or monetary policy eases (a lower ii schedule). The AD schedule relates aggregate demand, output and inflation. Next, we turn to aggregate supply.



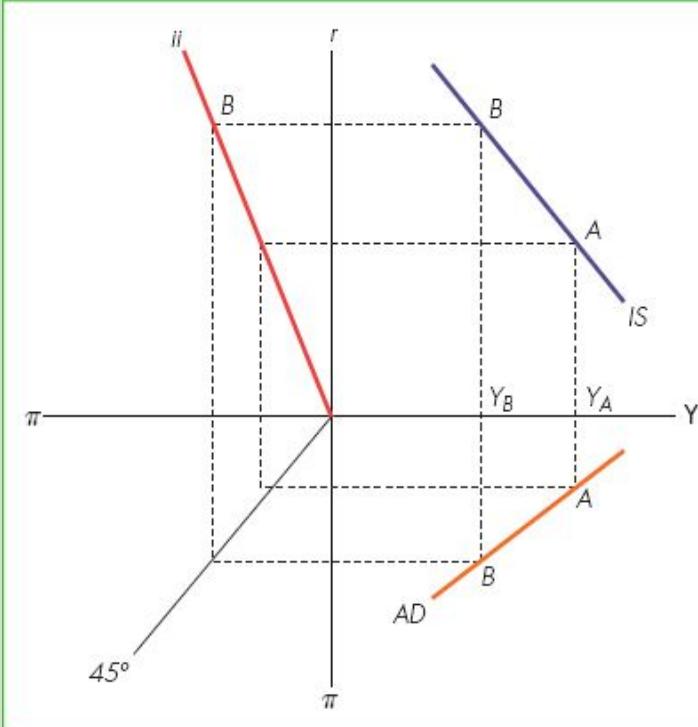
The aggregate demand schedule AD shows that higher inflation reduces aggregate demand by inducing the central bank to raise real interest rates.

Figure 21.2 The aggregate demand schedule

CONCEPT 21.2

AGGREGATE DEMAND, THE IS SCHEDULE AND THE II SCHEDULE

By now, any sensible person is asking two questions. First, why are there so many ways in which to model aggregate demand? Second, how do these different approaches fit together? The figure provides the answers.



We measure output on the horizontal axis in the right-hand panels – a movement to the right denotes higher output and aggregate demand. We measure nominal interest rates in the upward vertical direction – higher up denotes a higher nominal interest rate. We measure inflation in the horizontal direction in the left-hand panel – a movement further to the left denotes higher inflation. We also measure inflation in the bottom panel – a movement further down denotes higher inflation. The 45-degree line ensures that whatever inflation rate we measure in the horizontal direction is the same as the inflation rate we measure in the vertical direction.

Begin in the top right-hand panel, which relates aggregate demand to nominal interest rates. The *IS* schedule shows, other things equal, how lower interest rates boost aggregate demand as we move along the *IS* schedule. Any other change, such as a fiscal expansion, would be shown as an upward shift in the *IS* schedule.

The top left-hand panel relates nominal interest rates to inflation. The *ii* schedule describes the given monetary policy in force. As drawn, any particular increase in inflation leads to a larger increase in nominal interest rates – the central bank ensures that real interest rates rise when inflation rises in order to keep inflation under control.

Thus, if we begin at point *A* in the top right-hand panel, we can deduce the associated nominal interest rate on the vertical axis, and, for the given monetary policy, the associated inflation rate on the horizontal axis. Using the 45-degree line, we convert this to the same inflation rate on the vertical axis.

Hence we derive point *A* in the bottom right-hand panel. This shows the implied inflation rate when output and aggregate demand are Y_A .

Similarly, if we begin at point *B* in the top right-hand panel, we can again infer the corresponding levels of nominal interest rates and inflation, leading to point *B* in the bottom right-hand panel. Repeating this for each of the points on the *IS* schedule in the top right-hand panel, we trace out the aggregate demand schedule *AD* in the bottom right-hand panel; that is, the entire relationship between output demand and inflation.

Movements along the *AD* schedule show how demand, interest rates and inflation rates move together. What would shift the aggregate demand schedule? The 45-degree line cannot change, so the answer has to be a shift in the *ii* schedule (a change in monetary policy) or a shift in the *IS* schedule (autonomous demand changes caused by shifts in consumer demand, investment demand, government demand or export demand – everything except the effect of changes in interest rates).

To do everything properly, we could carry the entire toolkit of the above figure around. But it is quite a heavy toolbox. Often, we can get by using part of the analysis – just the *ii* schedule, just the *AD* schedule or just the *IS* schedule. However, if you ever find yourself struggling to understand the full picture, it is to the diagram above that you should eventually return. It will not let you down.

21.2 Aggregate supply

The **aggregate supply schedule** shows the output that firms wish to supply at each inflation rate.

At **potential output** all inputs are fully employed.

Long-run equilibrium output occurs when physical production inputs are fully employed. This output level is independent of inflation.

In the classical model, **the aggregate supply schedule** is vertical at potential output.

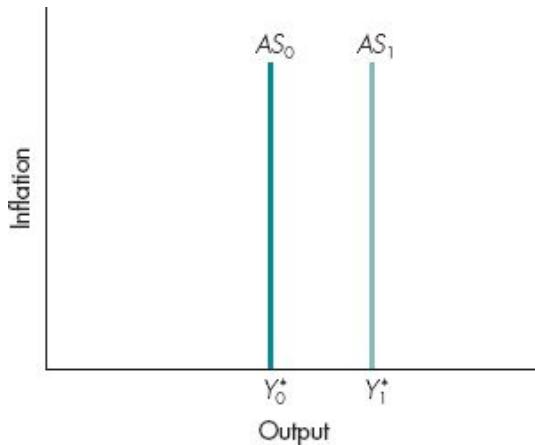
When prices and wages are completely flexible, output is always at **potential output**.

Potential output depends on the level of technology, the quantities of available inputs (labour, capital, land, energy) in long-run equilibrium, and the efficiency with which resources and technology are exploited. In the long run, investment in physical and human capital raises inputs of labour and capital, technical progress improves technology, and supply-side policies reduce distortions and raise efficiency. In the short run, we treat potential output as given; it is **long-run equilibrium output**.

With flexible wages and prices, how does a rise in inflation (and hence faster growth of nominal wages) affect the incentive of firms to supply goods and services?

Thinking in real terms, firms compare the real wage (the nominal wage W divided by the price level P) with the real benefit of labour, the extra output it makes. Similarly, workers compare real take-home pay (purchasing power over goods and services) with the disutility of sacrificing more leisure in order to work longer. If wages and prices both double, real wages are unaffected. Neither firms nor workers should change their behaviour. **Aggregate supply** is unaffected by pure inflation since everything nominal rises by the same proportion, as shown in Figure 21.3.

Money illusion exists if behaviour responds to changes in nominal variables rather than real variables.



In the classical model, aggregate supply equals potential output, whatever the inflation rate. The AS schedule is vertical. A rise in potential output, from Y_0^* to Y_1^* shifts aggregate supply from AS_0 to AS_1 .

Figure 21.3 The vertical AS schedule

Wage and price flexibility ensures all nominal variables rise together. Without **money illusion**, people see through nominal changes: real variables are unaltered.

In the classical model, real things determine real things, and nominal things determine other nominal things. Better technology, more capital or greater labour supply raise potential output, shifting the vertical supply curve from AS_0 to AS_1 in Figure 21.3. However, for any given level of potential output, lower inflation does *not* reduce the real output that firms wish to supply.

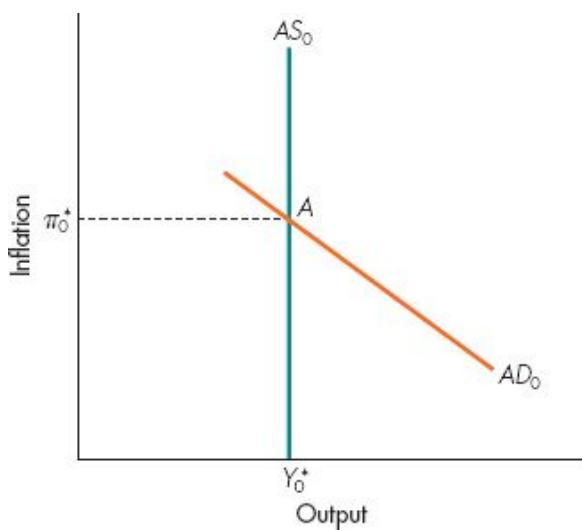
21.3 Equilibrium inflation

For the classical model, Figure 21.4 shows the aggregate demand schedule AD_0 and the vertical aggregate supply schedule AS_0 . Output is at potential output and inflation is π^*_0 . At point *A*, there is equilibrium in all markets: for output, money and labour.

The labour market is in equilibrium anywhere on the AS_0 schedule, since the economy is at potential output and full employment. *A* is also on the aggregate demand schedule along which interest rates are adjusted in line with monetary policy and the aggregate demand for goods equals the actual output of goods.

The equilibrium inflation rate π^*_0 reflects the positions of the AS and AD schedules. Potential output y^*_0 reflects technology, efficiency and available input supplies. The macroeconomic demand schedule depends on the IS schedule, showing how interest rates affect aggregate demand, and on the ii schedule of Figure 21.1, showing how interest rates respond to deviations of inflation from its target level.

To ensure that equilibrium inflation π^*_0 coincides with the inflation target π^* , the central bank chooses the correct height of the ii schedule in Figure 21.1, thereby ensuring the AD schedule has the correct height to make equilibrium inflation π^*_0 coincide with the target inflation rate π^* . If π^*_0 is too low, the central bank loosens monetary policy, shifting the ii schedule down and the AD schedule up. If π^*_0 exceeds the inflation target, a tighter monetary policy shifts the ii schedule up and the AD schedule down.



With aggregate supply AS_0 and aggregate demand AD_0 , inflation is π_0^* and output is Y_0^* .

Figure 21.4 Equilibrium inflation

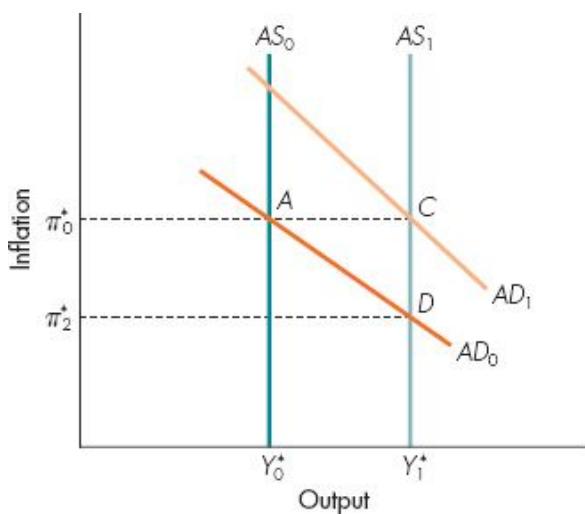
A supply shock

Monetary policy accommodates a permanent supply change by altering the real interest rate (shift in the ii schedule) to induce a similar change in aggregate demand.

Supply shocks may be beneficial, such as technical progress, or may be adverse, such as higher real oil prices or loss of capacity after an earthquake. Suppose potential output rises. In Figure 21.5 the AS schedule shifts to the right, from AS_0 to AS_1 . For a given AD schedule, equilibrium inflation falls to π_2^* , with equilibrium at D .

However, the central bank still wants a long-run equilibrium inflation rate π_0^* . Hence, in response to the supply shock, the central bank loosens monetary policy, shifting the ii schedule downwards and the AD schedule upwards. Lower real interest rates boost aggregate demand in line with higher potential output y_1^* . The new equilibrium is at C , not D . With unchanged inflation, the lower real interest rate also implies a lower nominal interest rate.

Lower interest rates raise the demand for money. To restore money market equilibrium, the central bank must then supply more money.



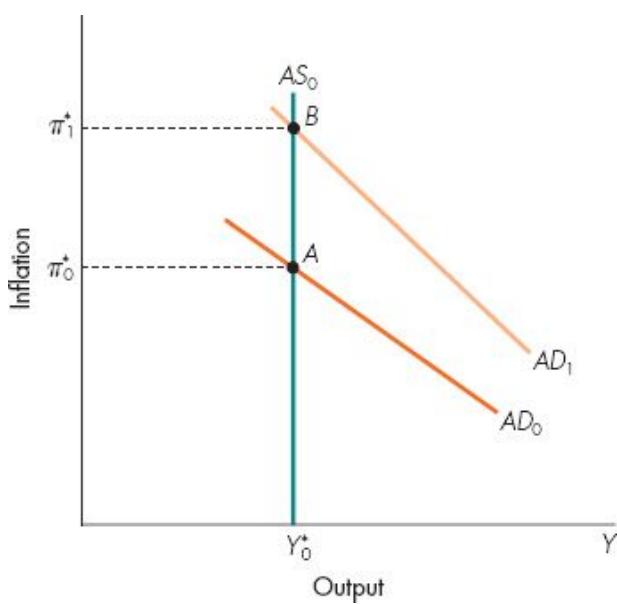
With aggregate supply AS_0 and aggregate demand AD_0 , inflation is π^* and output is Y_0^* . A rise in supply shifts aggregate supply from AS_0 to AS_1 . The central bank accommodates this extra supply, reducing i^* in order to shift demand to AD_1 , thus maintaining equilibrium inflation at π^* . Equilibrium then shifts from A to C.

Figure 21.5 A supply shock

Conversely, if high oil prices permanently reduce aggregate supply, this shifts AS_1 to AS_0 . Beginning at point C, the central bank must then tighten monetary policy so that higher real interest rates reduce aggregate demand in line with the lower aggregate supply.

A demand shock

Suppose aggregate demand shifts up because of easier fiscal policy or greater private sector optimism about future incomes and profits. Beginning from equilibrium at A in Figure 21.6, but keeping supply fixed at AS_0 , a demand shift from AD_0 to AD_1 leads to a new equilibrium at B.



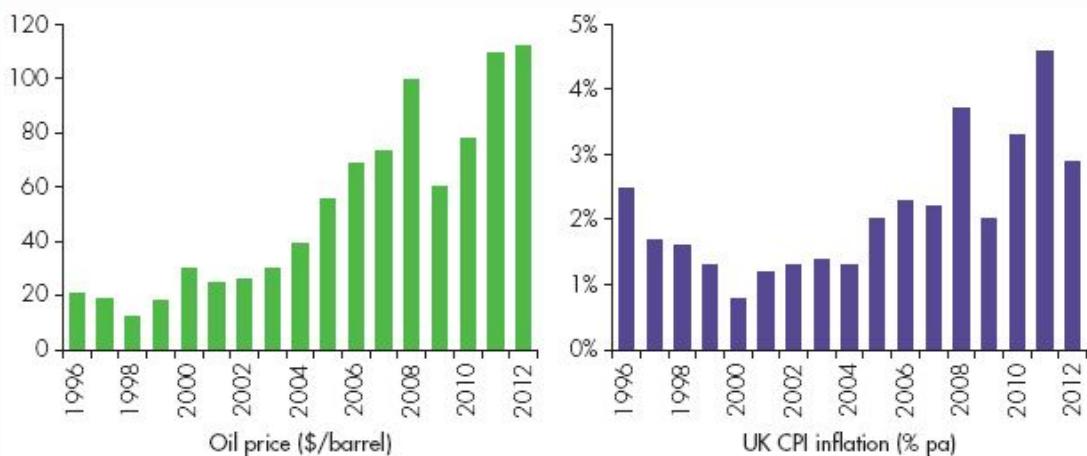
With aggregate supply AS_0 and aggregate demand AD_0 , inflation is π^*0 and output is Y^*0 . For a given aggregate supply, a rise in demand from AD_0 to AD_1 , violates the long-run inflation target at π^* . Thus the central bank raises i^* to shift AD_1 back to AD_0 and restore equilibrium at A.

Figure 21.6 A demand shock

CASE21.1

SUPPLY SHOCKS: OIL PRICES AND INFLATION

The figure below shows the dramatic increase in oil prices after 2003. If oil price shocks lead to inflation, why did so little inflation materialize? Was the Bank of England lulled into a false sense of security? Should we be surprised that, by April 2007, the Bank had to justify why it had allowed UK inflation to exceed the target range to which it is committed? Was it only the financial crash that spared the Bank further embarrassment?



The early years of inflation targeting were a benevolent environment for monetary policy. Globalization was flooding the West with cheap imports from China – helping to keep prices down – and trade unions in Europe and the US were aware that they could make domestic firms uncompetitive by pressing too hard for wage increases.

This same globalization put upward pressure not just on oil but on other commodity prices. Once China and India, the world's two most-populous economies, were growing at 8–10 per cent a year, their demands for raw materials were substantial: demand grew faster than supply could keep up, so commodity prices increased.

The tripling of oil prices during 2003–07 was potentially as dramatic a supply shock as the world had experienced when OPEC first flexed its muscles in 1973. Yet the effects were very different. In the 1970s, oil price rises led to a decade of high inflation and output stagnation. In the 2000s, economies continued to boom until the financial crash, and inflation remained muted.

Why was the impact of higher oil prices so modest? First, the increase took place steadily rather than in a step jump. Having longer to adjust, purchasers of oil had more time to switch to alternative energy sources. Second, the pressure of globalization encouraged importing countries to absorb the oil price increase rather than attempt to pass it on in higher prices. Workers accepted reductions in real wages, and firms accepted a reduction in profit margins.

Central banks argued that inflation targeting had affected the outcome in a helpful way. Firms and workers understood that attempts to pass on higher energy costs into higher wages and output prices would simply induce the central bank to raise interest rates dramatically causing a recession that would be more painful than absorbing the oil price rise. By anchoring inflation

expectations, monetary policy encouraged adjustment in real wages and real profit margins.

Even so, by 2008, inflation was clearly increasing sharply. We will never know what would have happened in the absence of the financial crash. Presumably world commodity prices would have remained high, and central banks would have had to raise interest rates, possibly quite a bit, to get inflation back on track.

The financial crash imposed a sharp fall in aggregate demand, and initiated rapid contraction. World demand and world commodity prices fell for a bit. But by 2009, China had returned to 10 per cent growth, and the price of oil and other commodities began rising again. Significant growth rates were being seen in many emerging markets. Upward pressure on world commodity prices was more intense than ever.

Despite the post-crash austerity experienced in Europe, Japan and even the US, these countries continued to face rising import prices, creating an acute dilemma for monetary policy: raise interest rates to choke off imported inflation, or leave interest rates low to support the real economy at a time when fiscally challenged governments had few other tools with which to stave off outright depression.

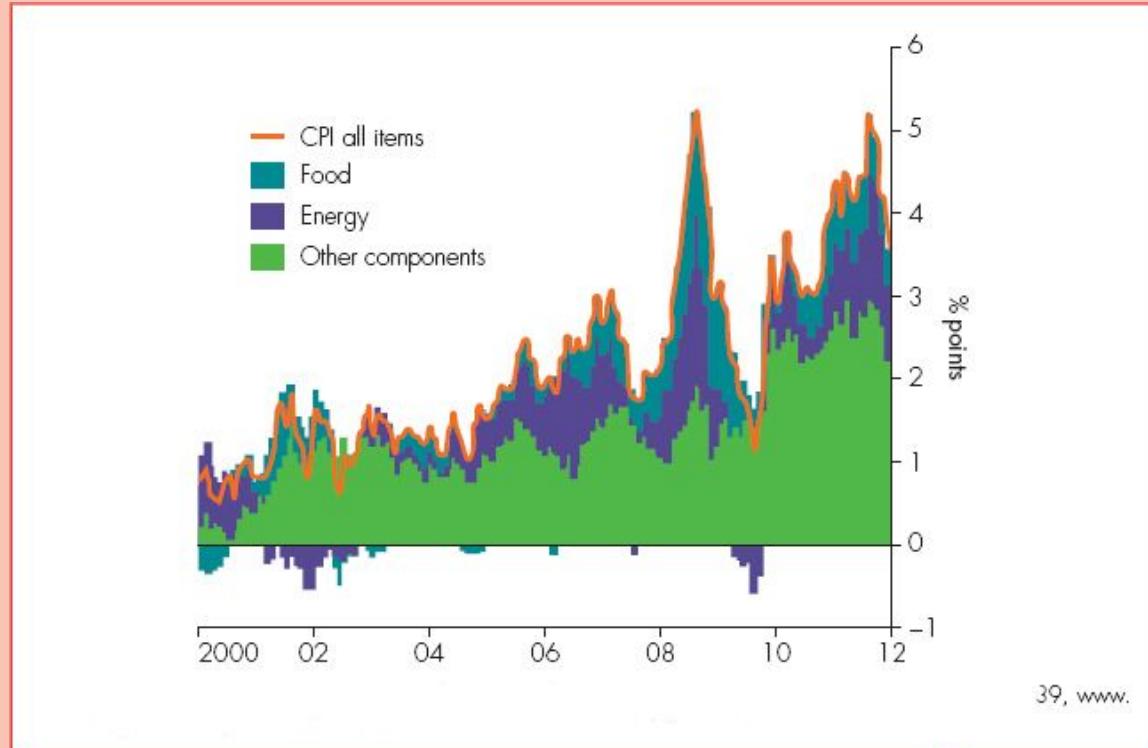
Central banks in Europe, the US and Japan opted to keep interest rates low even if this meant deviating from the inflation target for a while. The right-hand chart shows that UK inflation tumbled in 2009 in the post-crash austerity. However, UK inflation picked up again in 2010 because of tax rises and exchange rate depreciation.

The Bank of England announced in advance that it would regard the inevitable rise in UK inflation during 2010 as temporary, and would not expect to raise interest rates immediately to stave this off. Given the fragility of the economy, it preferred to keep interest rates low for a little longer. In contrast, the ECB temporarily raised interest rates in 2011 to combat the rising inflation faced in the Eurozone, preferring to prioritize low inflation over output recovery.

This episode illustrates the differing degrees of flexibility with which a credible central bank can pursue inflation targeting. The more people believe that the central bank will keep a grip on inflation in the long run, the less people will care about blips in the short run.

Taken from the Bank of England's *Quarterly Bulletin* 2012 Q1, the chart below summarizes all the different contributory factors to UK inflation since 2000. Adverse supply shocks contributed substantially to inflation during

2008–09 and 2011, but usually the turquoise and purple components account for much less than half of total inflation.



The central bank can continue to hit its inflation target π_0^* only by tightening monetary policy to offset the demand shock. In full equilibrium, with unchanged supply AS_0 , aggregate demand must not change. By raising real interest rates, the central bank can reduce aggregate demand again. The central bank thus tightens monetary policy (an upward shift in the ii schedule) until the demand shock is fully offset and AD_1 has shifted down to AD_0 again. Equilibrium remains at A and the inflation target π_0^* is still achieved.

The original rise in demand could have come from the private or the public sector. If from higher private demand, the higher real interest rate simply reduces private demand back to its original level. If from higher government spending, the central bank raises interest rates until private spending falls by as much as government spending increased.

In the classical model with a vertical AS schedule, a rise in government spending **crowds out** an equal amount of private spending. Aggregate demand remains equal to potential output.

Note the distinction between partial crowding out in the Keynesian model and complete [crowding out in the classical model](#). In the Keynesian model, output was demand-determined in the short run. Higher *output* induced the central bank to raise interest rates, which partly offset the expansionary effect of higher government spending. In the classical model, aggregate supply is the binding constraint. Output does not change. When higher government expenditure raises aggregate demand, higher interest rates must reduce consumption and investment to leave aggregate demand unaltered.

We may draw a second conclusion from Figure 21.6. Suppose monetary policy changes because the inflation target is raised from π^*_0 to π^*_i . With a higher target inflation rate, the central bank no longer needs such high real interest rates at any particular level of inflation. Real interest rates fall and the aggregate demand schedule shifts up from AD_0 to AD_1 . With an unchanged *AS* schedule, equilibrium moves from *A* to *B*.

In the new equilibrium, inflation is higher but real output is unaltered. Since it is a full equilibrium, all real variables are then constant. One of these variables is the real money stock M/P . Since prices grow at the rate π^*_i , the nominal money supply must also grow at this rate.

In the classical model, faster [nominal money growth](#) is accompanied by higher inflation but leaves real output constant at potential output.

The idea that [nominal money growth](#) is associated with inflation, but not growth of output or employment, is the central tenet of *monetarists*. Figure 21.6 shows this is correct in the classical model with full wage and price flexibility and no money illusion.

How long does all this take?

The classical model studies the economy once all variables have fully adjusted. Instead of thinking of adjustment as instant, we can view the classical model as applying to a long enough time for slower adjustment to be completed. This means not just wage and price adjustment, but also time for the central bank to work out what is going on and amend monetary policy if necessary, and time for these interest rate changes to have their full effect on private behaviour. Suppose the economy faces a fall in aggregate demand. What happens next?

The classical model

With aggregate supply unaffected, a fall in aggregate demand leads to lower inflation, to which the central bank immediately responds by easing monetary policy, reducing the real interest rate, boosting private sector demand and thus restoring aggregate demand to the unchanged level of potential output.³

The Keynesian model

Before wage and price adjustment is possible, there is no change in inflation to which the central bank can respond. The initial effect of lower aggregate demand is simply a fall in output. The rest of this chapter studies the adjustment process by which the economy gradually makes the transition from the Keynesian short run to the classical long run. To do so, we introduce the short-run aggregate supply curve.

21.4 The labour market and wage behaviour

Downward shocks cause recessions lasting years not weeks. Why don't changes in prices react faster, allowing changes that restore potential output? Firms relate prices to costs. Wages are the largest part of costs. Sluggish wage adjustment to departures from full employment is the main cause of slow adjustment of prices.

For both firms and workers, a job is often a long-term commitment. For the firm, it is costly to hire and fire workers. Firing entails a redundancy payment and the loss of the expertise the worker had built up on the job. Hiring entails advertising, interviewing and training a new worker in the special features of that firm. Firms are reluctant to hire and fire workers just because of short-term fluctuations in demand.

For the worker, seeking a new job takes time and effort, and throws away experience, seniority and the high wages justified by the high productivity that comes from having mastered a particular job in a particular firm. Like firms, workers care about long-term arrangements. Firms and workers reach an understanding about pay and conditions *in the medium term*, including how to handle fluctuations in the firm's output in the short run.

A firm and its workers have explicit contracts, or implicit agreements, specifying working conditions. These include normal hours, overtime requirements, regular wages and pay schedules for overtime work. The firm then sets the number of hours, within the limits of these conditions, depending on how much output it wishes to make in that week.

When demand falls, the firm initially reduces hours of work. Overtime ends and factories close early. If demand does not recover, or declines further, firms start firing workers. Conversely, in a boom a firm makes its existing workforce work overtime. Then it seeks temporary workers to supplement the existing labour force. Only when the firm is sure that higher sales will be sustained does it hire extra permanent workers.

Wage adjustment

Wages are not set in a daily auction in which the equilibrium wage clears the market for labour. Firms and workers both gain from long-term understandings.

This mutual commitment partly insulates a firm and its workforce from temporary conditions in the labour market.

Nor can a firm and its workforce spend every day haggling. Bargaining is costly, using up valuable time that could be used to produce output. Bargaining costs mean wages change only at discrete intervals. Immediate wage adjustment to shocks is ruled out. At best, firms must wait until the next scheduled date for a revision in the wage structure. In practice, complete wage adjustment is unlikely to take place even then. Chapter 10 discussed other reasons why involuntary unemployment is not instantly eliminated by wage adjustment.

Recap

In the short run (first few months), changes in labour input are largely changes in hours. In the medium run (up to two years), as changes in labour demand persist, the firm begins to alter its permanent workforce. In the long run (perhaps four to six years), adjustment is complete.

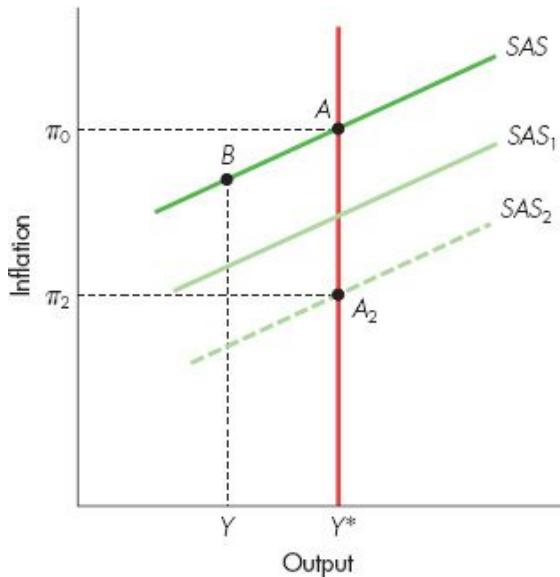
In the short run, trends in wages are largely given. The firm can affect earnings, as distinct from negotiated wage rates, because fluctuations in overtime and short time affect average hourly earnings. This flexibility is limited. In the medium run, the firm begins to adjust the path of wages. In the long run, the process is complete and the economy is back at potential output.

Now think about the market for output. By distinguishing between supply in the short and long run, our model of output reflects *both* supply and demand, even in the short run. Nevertheless, its short-run behaviour is like the simple Keynesian case in which output is demand-determined. Its long-run behaviour is fully classical.

21.5 Short-run aggregate supply

In Figure 21.7 the economy is at potential output at A . In the short run, the firm inherits a given rate of nominal wage growth (not shown in the figure). Previous wage negotiations anticipated remaining in long-run equilibrium at A with inflation p_0 . By keeping up with inflation, nominal wage growth is expected to maintain the correct real wage for labour market equilibrium.

If inflation exceeds the expected inflation rate p_0 , this helps firms by raising their output prices. The real wage is lower than expected. If this had been foreseen when wages were negotiated, the inherited nominal wage would have been higher; but it was not foreseen. Firms take advantage of their good luck by supplying a lot more output. They can afford to pay overtime to ensure that the workforce co-operates, and may also take on temporary extra staff.



Firms raise prices when wage costs rise. Each short-run aggregate supply schedule reflects a different rate of inherited nominal wage growth. For any given rate, higher inflation moves firms up a given short-run supply schedule. A persisting boom or slump gradually bids nominal wage growth up or down, shifting short-run aggregate supply schedules. When these shift enough to restore to the inflation rate at which AD and AS intersect, potential output is restored.

Figure 21.7 Short-run aggregate supply

The **short-run supply curve(SAS)** shows how desired output varies with inflation, for a given inherited growth of nominal wages.

Conversely, if inflation is below p_0 , the real wage is now higher than anticipated when the nominal wage was agreed. Since labour is now costly, firms cut back output a lot. They move from A to B in Figure 21.7. Firms move along the **short-run supply curve (SAS)** in the short run.

If demand and output remain low, the growth rate of negotiated nominal wages gradually falls. With lower wage growth, firms do not need to raise output prices so quickly. The short-run aggregate supply schedule shifts down from SAS to SAS_1 in Figure 21.7. Lower inflation moves the economy down its aggregate demand schedule, increasing the demand for goods. If full employment and potential output are still not restored, negotiated wage growth falls again, leading to a short-run aggregate supply schedule such as SAS_2 .

These short-run aggregate supply schedules give a realistic picture of adjustment to demand shocks. Because the short-run aggregate supply schedule is flat, a shift in aggregate demand leads mainly to changes in output not prices in the short run. This is the Keynesian feature. But deviations from full employment gradually change wage growth and short-run aggregate supply.

The economy gradually works its way back to potential output. That is the classical feature. We now describe adjustment in more detail.

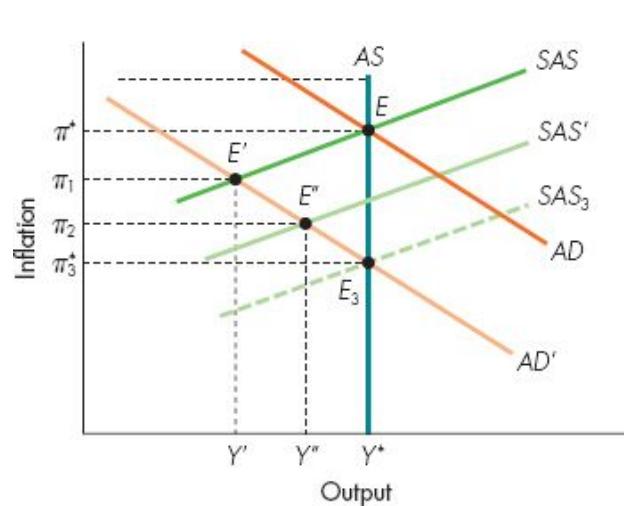
21.6 The adjustment process

We now combine the aggregate demand schedule with the short-run aggregate supply schedule to show how demand or supply shocks set up an adjustment process. In so doing, we now assume that the goods market clears, even in the short run. Short-run aggregate supply gradually changes over time as wage growth adjusts to the rate that restores full employment and potential output, placing firms eventually on their long-run aggregate supply schedule.

Output is no longer demand-determined when aggregate demand lies below the level of potential output. In the short run, firms are also on their short-run supply schedules producing what they wish, *given the inherited nominal wages*.

However, sluggish wage adjustment prevents immediate restoration of full employment. When aggregate demand for goods falls, firms reduce output and employment. Since wages do not fall at once, there is involuntary unemployment. *Employment* is demand-determined in the short run.

Figure 21.8 shows a downward shift in the aggregate demand schedule from AD to AD' because monetary policy is tightened (a higher ii schedule in Figure 21.1). In the long run, aggregate demand must return to potential output, and the economy will end up at E_3 . Hence, the tighter monetary policy can be viewed as a cut in the target inflation rate from π^* to π_3^*



Beginning at E , a lower inflation target shifts AD to AD' . Given inherited wage growth, the new equilibrium is at E' . Output falls from Y^* to Y' , and actual inflation is only π_1 . Since wages have risen faster than prices despite the fall in output, unemployment rises. In the next wage settlement, nominal wage growth slows, and the short-run supply schedule becomes SAS' . Equilibrium is now at E'' , and output recovers to Y'' . Once wage growth slows enough to make SAS_3 the supply curve, long-run equilibrium is re-established at E_3 .

Figure 21.8 A lower inflation target

When monetary policy is tightened, interest rates initially rise since actual inflation at E is now above target. Aggregate demand shifts down to AD' . In the classical model, there is an instant adjustment of prices and wages to keep the economy at full employment and potential output. Equilibrium inflation immediately falls to π_3^* and the new equilibrium is at E_3 . Output remains at potential output Y^* .

These classical results are valid only in the long run. When adjustment of wages and prices is slow, the economy faces the short-run aggregate supply schedule SAS , reflecting the nominal wages recently agreed.

In the short run, the downward shift in AD causes a move from E to E' . Since firms cannot cut costs much, they reduce output to Y' . At E' the goods market clears at the intersection of the aggregate demand schedule AD' and the supply schedule SAS . Inflation has fallen a little because of lower demand, but output has fallen a lot. With lower inflation than the expectation built into nominal wage agreements, *real wages have risen*, despite the fall in output. Once firms can adjust employment, some workers are fired and unemployment rises.

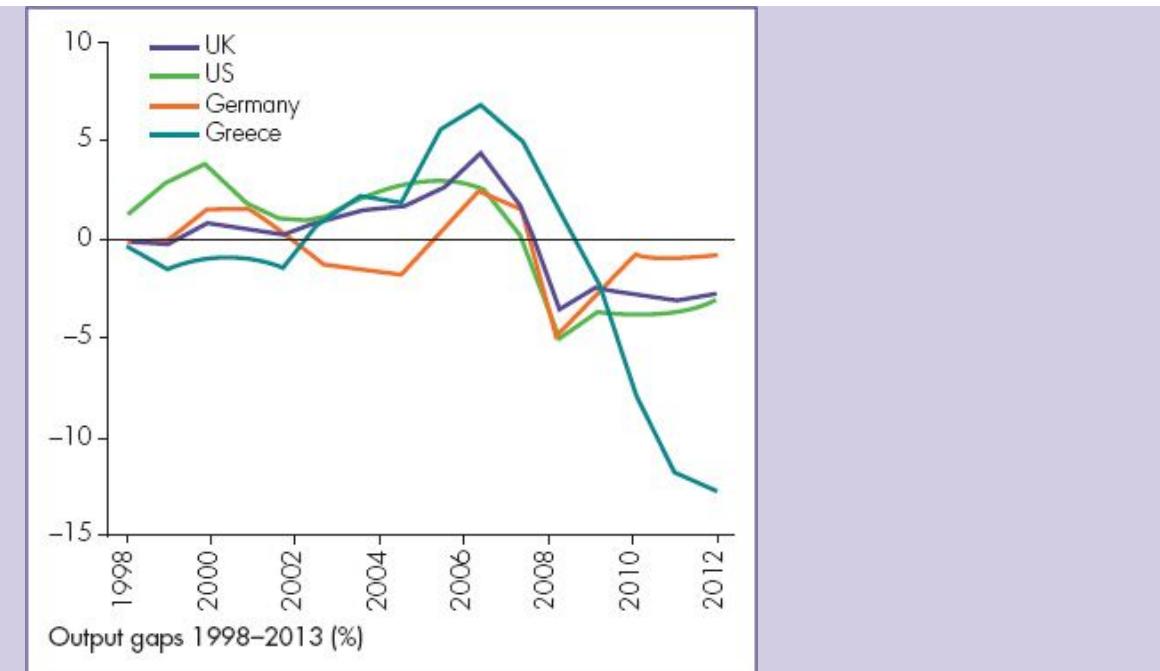
In the medium run, this starts to reduce wage growth. With inherited wages lower than they would have been, firms move on to a lower short-run aggregate supply schedule SAS' . The goods market now clears at E'' . Output and employment recover a bit, but some unemployment persists. Since inflation has fallen, the central bank is less worried about the amount by which inflation exceeds its new target and cuts real interest rates, moving the economy down AD' to E_0 .

In the long run, adjustment is complete. Wage growth and inflation fall to π_3^* . The short-run aggregate supply schedule is SAS_3 in Figure 21.8. The economy is in full equilibrium at E_3 , on AS , SAS_3 and AD' . Output is Y^* and the labour market is back at full employment.

The real world lies between the extreme simplifications of the simple Keynesian model and classical models. In practice, prices and wages are neither fully flexible nor fully fixed. A tougher inflation target has real effects in the short run, since output and employment are reduced. But after wages and prices adjust fully, output and employment return to normal. Inflation is permanently lower thereafter.

ACTIVITY 21.1

OUTPUT GAPS, 1998–2013



The output gap ($Y - Y^*$) is the percentage deviation of actual output Y from potential output Y^* . Each

year the OECD *Economic Outlook* estimates potential output for all its member countries. The diagram shows estimates for the UK, the US and Germany. Positive output gaps are booms; negative gaps indicate slumps.

The diagram shows the relative stability of the period 1998–2006. Central banks were successfully managing aggregate demand to keep it close to full capacity.

The diagram also shows that all four countries were overheating in 2007/08. Aggregate demand was higher than potential output, and in the UK, Greece and Germany had increased sharply since 2005. Greece was experiencing a boom that was completely unsustainable, with demand almost 6 per cent in excess of potential output.

When the financial crisis hit, some economists thought that Germany would be relatively well insulated, since its regulation of banks had been more stringent than in the UK and the US. Yet, the diagram shows that Germany experienced nearly as dramatic a slump in aggregate demand as its Anglo-Saxon competitors. Sub-prime mortgages had found their way even into Stuttgart and Frankfurt. Even China did not escape. When aggregate demand in China fell in 2009, German exports were hard hit. Thus different countries experienced the crash through different channels. The crisis originated in the US, and UK banks were then very exposed. Germany suffered both because all export markets suffered, and because its banks were exposed to banks in other countries that themselves had behaved less prudently.

Nowhere was the turnaround more acute than in Greece. Over the six-year period from 2007 to 2013, aggregate demand for Greek output fell by more than 20 per cent. In part, this reflected the same world recession and banking crises experienced by other countries, but it was exacerbated by the need to seek several bailouts from foreign governments, whose terms for emergency assistance included additional fiscal austerity in Greece.

Greece aside, the diagram shows that the upturn is slowly starting to take effect. It also confirms that for the next few years all major economies will have substantial spare capacity – the underlying assumption of the Keynesian perspective.

Finally, the diagram helps identify periods in which simple Keynesian analysis cannot be the whole story. Once the output gap has been eliminated, as during the period around 2000, there is no spare capacity remaining, and the classical model is increasingly relevant.

Questions

There are two ways in which you might try to calculate potential output, and hence the output gap: (i) statistically, by fitting trend lines through previous business cycles or (ii) economically, by trying to get an idea of the balance of aggregate supply and aggregate demand.

- a. If you wanted a quick procedure capable of being replicated across many countries, which of the two would you be inclined to choose?
- b. How might you build up an idea of an empirical economic model of the balance between actual output and potential output?
- c. A central bank reduces interest rates but is disappointed to find that this quickly generates higher inflation not higher output. What can you infer about the initial level of the output gap? Why?

To check your answers to these questions, go to page 680.

21.7 Sluggish adjustment to shocks

A permanent supply shock

Suppose a change in attitudes towards women working leads to an increase in labour supply. Potential output rises. In the long run, aggregate demand must rise in line with aggregate supply. Lower real interest rates allow higher aggregate demand at the unchanged inflation target π^* . Provided monetary policy is loosened, the rightward shift in AD can match the rightward shift in aggregate supply. By accommodating the extra supply with looser monetary policy, the

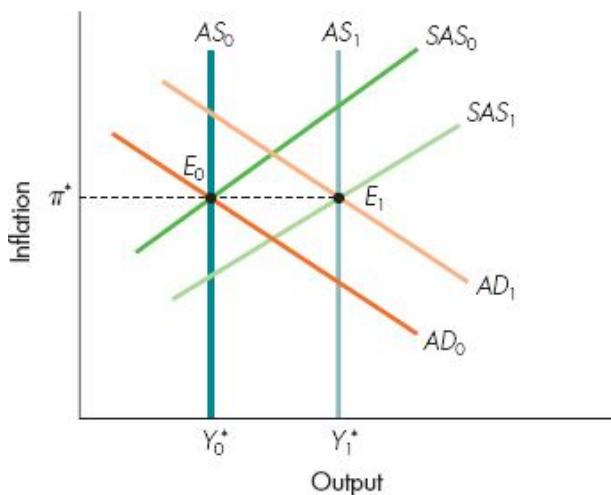
inflation rate remains π^* , and the economy moves directly to the new long-run equilibrium, from E_0 to E_1 in Figure 21.9.

A permanent supply shock changes the level of potential output.

Because of lags in diagnosing the shock, and in the response of consumption and investment demand to lower interest rates, Figure 21.9 exaggerates the ease of adjustment to a permanent supply shock. In practice, output may not jump all the way to the new level of potential output.

If the aggregate demand schedule does not fully and immediately shift to AD_1 , output is below π^* . This reduces inflation and the central bank responds with lower interest rates. Over time, the aggregate demand schedule will drift to the right until it reaches AD_1 in Figure 21.9.

A temporary supply shock



A permanent rise in supply shifts AS_0 and SAS_0 to AS_1 and SAS_1 . By permanently reducing interest rates, the central bank shifts AD_0 to AD_1 , meeting its inflation target π^* in the new equilibrium at E_1 . If the central bank acts quickly, no further shifts in SAS_1 are required.

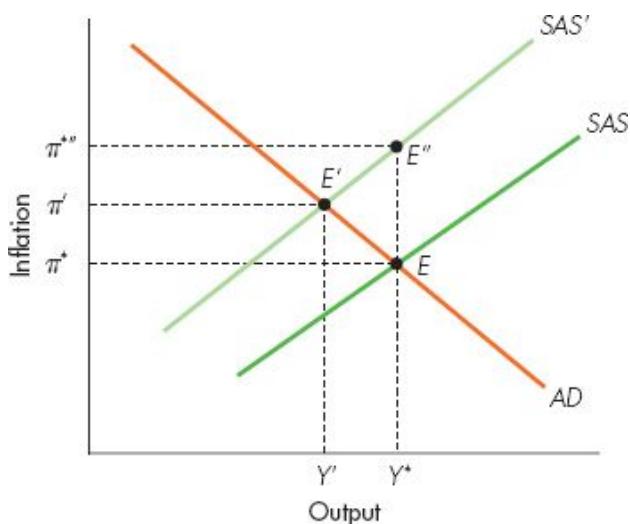
Figure 21.9 A permanent supply increase

A temporary supply shock leaves potential output unaffected in the long run. With the vertical AS schedule unaltered, the short-run supply curve must shift. Although the SAS schedule is mainly influenced by inherited nominal wages, it is also affected by other input prices. Suppose a temporary oil price rise makes firms charge higher prices at any output level. Figure 21.10 shows a shift upwards in short-run supply, from SAS to SAS' . The new short-run equilibrium is at E' .

Inflation rises but output and employment fall because the central bank raises real interest rates in response to higher inflation.

If the central bank maintains its inflation target π^* , lower output and employment at E' gradually reduce inflation and nominal wage growth, shifting SAS' gradually back to SAS . The economy slowly moves down the AD schedule back to the original equilibrium at E'' .

A **temporary supply shock** shifts the short-run aggregate supply schedule, but leaves potential output unaltered.

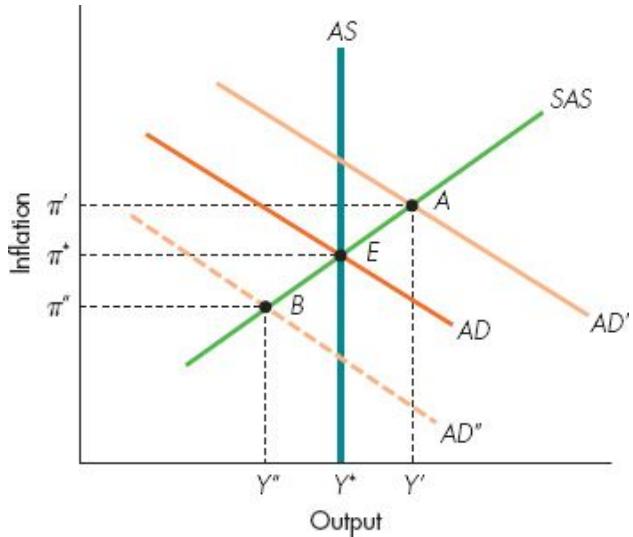


Higher oil prices force firms to raise prices. In the short run, SAS shifts up to SAS' , and equilibrium shifts from E to E' . Higher inflation reduces aggregate demand since the central bank raises real interest rates. Once the temporary supply shock disappears, SAS' gradually falls back to SAS , and equilibrium is eventually restored at E .

Figure 21.10 A permanent supply increase

A different outcome is possible. When the higher oil price shifts SAS to SAS_9 , it is possible to *avoid* the period of low output as the economy moves along AD from E_9 back to E . A *change* in monetary policy can *shift* AD up enough to pass through E_9 . Output can quickly return to potential output, but only because the inflation target⁴ has been loosened from π^{**} . The new long-run new equilibrium is then at E'' .

Monetary policy accommodates a temporary supply shock when monetary policy is altered to help stabilize output. The consequence, however, is higher inflation.



Demand fluctuates between AD' and AD'' , causing fluctuations in output and inflation. If the central bank can react quickly, it can offset demand shocks by changing i^* to shift demand back to AD . Stabilizing inflation at π^* has the effect of stabilizing output at Y^* .

Figure 21.11 Demand shocks

A central bank caring a lot about output stability may **accommodate short-run supply shocks**, even if this means higher inflation. A central bank caring more about its inflation target than about output stability will not accommodate temporary supply shocks.

It matters a lot whether the supply shock is temporary or permanent. If potential output is *permanently* affected, aggregate demand *must* eventually rise to match. Once a supply side shock is diagnosed as permanent, it should be accommodated.

Demand shocks

Figure 21.11 explores demand shocks *not* caused by monetary policy. If demand is high, facing AD' the economy moves along its short-run supply curve to point A . If demand is low, facing AD'' the economy moves along the SAS curve to point B .

Suppose the central bank diagnoses that an expansionary demand shock has occurred. It can tighten monetary policy and shift AD' back down to AD again. Similarly, it can loosen monetary policy in response to low aggregate demand AD'' , restoring AD again. The economy remains at E . Both inflation *and* output are stabilized.

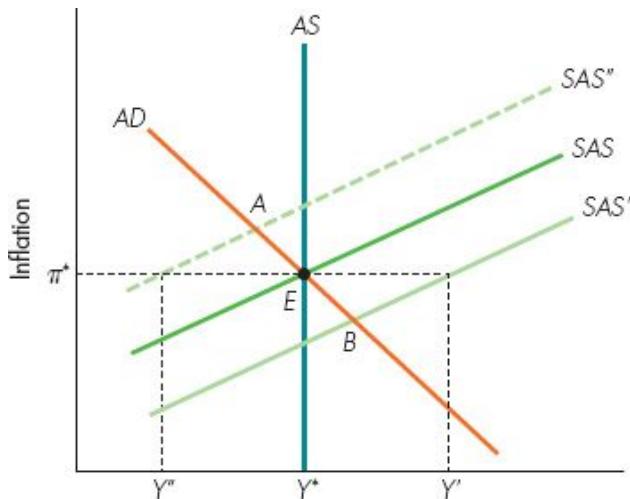
It is easy for the central bank to tell where inflation is relative to its target rate. It is harder to estimate the level of potential output, which can change over time. This is part of the modern case for using inflation targeting as the intermediate target of monetary policy. When all shocks are **demand shocks**, it works perfectly.

Suppose, instead, that all shocks are supply shocks. Figure 21.12 shows the long-run supply curve AS , vertical at potential output Y^* , and a set of short-run supply curves whose average level is SAS but which fluctuate between SAS_0 and SAS'' .

When all shocks are **demand shocks**, stabilizing inflation also stabilizes output, even in a Keynesian model.

On average, output is Y^* and inflation is π^* . If interest rates are varied very aggressively to stabilize inflation in the face of supply shocks, the AD schedule is effectively horizontal at π^* . Inflation is stabilized, but output fluctuates between Y'' and Y' when supply fluctuates between SAS' and SAS'' . Unlike the case of demand shocks, it is no longer possible to stabilize output *and* inflation.

Similarly, it is possible to stabilize output completely but only at the cost of allowing big fluctuations in inflation. The AD schedule is then vertical at potential output. A rise in short-run supply to SAS' induces a big rise in interest rates to reduce aggregate demand to Y^* again. With high supply but low demand, inflation is temporarily low (relative to inherited wage growth) and firms wish to supply only Y^* . When supply shrinks temporarily to SAS_0 , firms supply output Y^* only if inflation is high (relative to inherited wage growth), which needs a low interest rate to boost demand.



Short-run supply fluctuates between SAS' and SAS'' . If interest rates are set to stabilize inflation at π^* , output fluctuates between Y'' and Y' . Monetary policy cannot stabilize both output and inflation in response to supply shocks. It makes sense to set interest rates to allow some inflation fluctuations in order to reduce output a bit. The economy therefore fluctuates between points A and B.

Figure 21.12 Temporary supply shocks

21.8 Temporary supply shocks

Facing supply shocks, Figure 21.12 implies that it is a bad idea either to stabilize inflation at π^* (which induces big fluctuations in output) or to stabilize output at Y^* (which induces big fluctuations in inflation). The aggregate demand schedule AD in Figure 21.12 is a particular compromise in the way interest rates are set. The formula for the Taylor rule.⁵

Any AD schedule through point E achieves, on average, the targets π^* and Y^* . The particular schedule AD in Figure 21.12 makes the economy fluctuate between A (when supply is $SAS0$) and B (when supply is $SAS9$). This achieves acceptable fluctuations in both output and inflation. A steeper AD schedule, still through A , induces lower output fluctuations but larger inflation fluctuations. A flatter schedule has the opposite effect. The steepness of the schedule reflects the relative weight the central bank places on stabilizing inflation and output.

This trade-off does not arise for demand shocks. Figure 21.11 showed that, by fully offsetting demand shocks, the central bank stabilizes both output and prices. In reality, the central bank faces both supply and demand shocks, and cannot always diagnose which is which. It must choose a monetary policy that gives reasonable answers under both kinds of shock.

There is no conflict between output stability and inflation stability when shocks are demand shocks. It makes sense to try to hit the target as quickly as possible. Similarly, a permanent supply shock requires a permanent change in demand, which there is little reason to postpone. However, facing a *temporary supply shock*, Figure 21.12 showed that it makes sense temporarily to allow inflation to deviate from its target in order to mitigate the shock to output.

Flexible inflation targeting commits a central bank to hit inflation targets in the medium run, but gives it some discretion about how quickly to hit its inflation target.

The ii schedule in Figure 21.12 reflects the average behaviour of the central bank under flexible inflation targeting. Deviations of inflation from target are not all immediately eliminated, but they are eventually eliminated by the policy of raising (lowering) real interest rates whenever inflation is too high (low). Temporary deviations of inflation from target are the price to be paid for ensuring that output fluctuations are not too large. The key to successful **flexible inflation targeting** is that any deviation of inflation from target should be *temporary*.

CONCEPT 21.2

A TAYLOR RULE FOR MONETARY POLICY

Stanford professor John Taylor suggested that a neat way to describe flexible inflation targeting is to say that interest rates respond to deviations of both inflation and output from their target long-run equilibrium levels. Inflation above target, or output above target, is a signal to raise interest rates; inflation below target or output below target is a signal to reduce interest rates. We can think of the Taylor rule as applying to either nominal or real interest rates. However, a key insight of the Taylor rule is that, if inflation rises by 1 per cent, nominal interest rates must be increased by more than 1 per cent to ensure that real interest rates rise when inflation is higher. In the short run, the central bank – which must deal with a world in which both supply and demand shocks occur, and may not immediately be able to diagnose which – sets higher real interest rates if inflation exceeds the target and/or if real output exceeds potential output. Taylor showed that this formula provides a good empirical description of the behaviour of all major central banks.

In terms of rules versus discretion, this behaviour is not imposed on the central bank. It is largely the result of its discretionary behaviour, so it is a rule only in the sense of being a stable empirical relationship. But it does not reflect complete discretion. The target inflation rate itself is usually delegated by the government, not freely chosen by the central bank. For example, in the UK the government reserves the right to alter the inflation target during a crisis.

The Taylor rule also provides a way to indicate how rare and extreme an event the financial crash has been. For example, Princeton professor Paul Krugman, himself a Nobel Prize winner, took empirical estimates of the Taylor rule for the US and calculated that, at the height of the recession in 2009, ‘normal’ behaviour of the US central bank, given observed levels of inflation and output, would have implied a nominal interest rate of *minus* 5.6 per cent.⁶ Central banks would ideally liked to have cut interest rates way below zero if following their normal behaviour. The fact that they could not drive nominal interest rates below zero was really getting in the way of normal monetary policy.

This had two implications. First, fiscal policy was going to be asked to do ‘too much’ because monetary policy could not do enough. Second, quantitative easing was adopted not merely because of the need to offset the collapse of the bank deposit multiplier but also in the hope that the credit channel of monetary policy could help boost aggregate demand at a time when further interest rate cuts were not possible.

Taylor rule or nominal GDP target?

A Taylor rule allows the central bank to respond to changes in both real income and the price level. A special case of the Taylor rule is when the central bank sets a target for the path of nominal GDP. Since nominal GDP is

simply $P \cdot Y$, the price level multiplied by real output, a nominal income target is a Taylor rule in which prices and real output have *equal* weights in determining the central bank's response to deviations from target. A Taylor rule is more general since there is no requirement to place the same priority on inflation deviations and on real output deviations from target. In 2012, shortly after being confirmed as the next governor of the Bank of England, Mark Carney raised the possibility that the Bank of England should examine nominal GDP as a target for interest rate setting.

MATHS 21.1

THE FORMULA FOR THE TAYLOR rule

In the long run, the real interest rate is i^* , inflation is π^* and real output is Y^* . Formally, the Taylor rule implies that real interest i obeys

$$i - i^* = a(\pi - \pi^*) + b(Y - Y^*) \quad a > 0, b > 0 \quad (1)$$

The nominal interest rate r is simply the real interest rate i plus the inflation rate π ,

$$r - r^* = (1 + a)(\pi - \pi^*) + b(Y - Y^*) \quad a > 0, b > 0 \quad (2)$$

where $r^* = i^* + \pi^*$

Hence, the long-run target for nominal interest rates depends both on the long-run inflation target and on the desired level of real interest rates in the long run, which may depend, among other things, on the monetary–fiscal mix.

Thus, we can think of the Taylor rule as applying to either nominal or real interest rates, with the key requirement that the nominal interest version in equation (2) insists that any increase in inflation leads to a larger increase in nominal interest rates – by $(1 + a)$ times – in order to ensure that real interest rates move in the right direction to stabilize inflation. The absolute size of the parameters a and b tells us how aggressively monetary policy attempts to stabilize inflation and output. The relative size of the parameters tells us the relative importance of inflation and output to policy makers in the short run.

Many economists have noted that deviations of output from target are an important indicator of future inflation. Hence it is also possible to interpret the empirical success of the Taylor rule as implying central bank concern for current and future inflation, rather than for current output and current inflation.

Finally, as noted in Concept 21.2, empirical estimates of Taylor rules using data for the previous decade would have led to choices of nominal interest rates that were negative, which is not possible in practice. With no danger of inflation and output well below target, central banks would have loved to set negative interest rates if only they could. This led IMF chief economist Olivier Blanchard to note that a temporarily *higher* inflation target might have been one way in which real interest rates could become more negative. Suppose, for example, inflation was 4 per cent and nominal interest rates were still close to zero. Real interest rates would then have been 24 per cent, a powerful stimulus to aggregate demand. Mathematically, the Blanchard proposal achieves the appropriately negative left-hand side of equation (2) not by reducing r below zero but by raising π^* and hence r^* above previous levels. Most central banks were unenthusiastic about this proposal: whatever its short-run attraction, they feared it would then be hard to restore belief in a low inflation target again after the crisis was over.

Summary

- The **classical model** of macroeconomics assumes full flexibility of wages and prices and no money illusion.
- The *ii schedule* shows, under a policy of **inflation targeting**, how the central bank achieves high interest rates when inflation is high and low interest rates when inflation is low. Central banks set nominal not real interest rates, and hence must first forecast inflation in order to calculate what nominal interest rate they wish to set.
- The *ii* schedule shifts to the left, a higher real interest rate at each inflation rate, when monetary policy is tightened, and to the right, a lower real interest rate at each inflation rate, when monetary policy is loosened.
- The **aggregate demand schedule** shows how higher inflation reduces aggregate demand by inducing monetary policy to raise real interest rates.
- The classical model always has full employment. The **aggregate supply schedule** is vertical at **potential output**. **Equilibrium inflation** is at the intersection of the aggregate supply schedule and the aggregate demand schedule. The markets for goods, money and labour are all in equilibrium. Monetary policy is set to make the equilibrium inflation rate coincide with the inflation target.

- In the classical model, fiscal expansion cannot increase output. To continue to hit its inflation target, the central bank must raise real interest rates to restore aggregate demand to the level of potential output. **Higher government spending crowds out an equal amount of private spending**, leaving demand and output unaltered.
- Changing the target inflation rate leads to an equivalent change in the growth of wages and nominal money in the classical model, but not to a change in output.
- In practice, wages adjust slowly to shocks since job arrangements are long term. **Wage adjustment** is sluggish not merely because wage bargaining is infrequent, but also because workers prefer their long-term employers to smooth wages.
- Prices reflect mainly labour costs. The **short-run aggregate supply schedule** shows firms' desired output, given the inherited growth of nominal wages. Output is temporarily responsive to inflation, since nominal wages are already determined. As wage adjustment occurs, the short-run supply schedule shifts.
- The **Keynesian model** is a good guide to short-term behaviour but the **classical model** describes behaviour in the long run.
- **Permanent supply shocks** alter potential output. **Temporary supply shocks** merely alter the short-run supply curve for a while.
- If its effects were instant, monetary policy could completely offset **demand shocks**, stabilizing both inflation and output. **Temporary supply shocks** force a trade-off between output stability and inflation stability. The output effect of **permanent supply shocks** cannot be escaped indefinitely.
- **Flexible inflation targeting** implies the central bank need not immediately hit its inflation target, allowing some scope for temporary action to cushion output fluctuations.
- A **Taylor rule** reviews interest rate decisions as responding to both deviations of output from target and deviations of inflation from target. Except during the financial crash, when interest rates could not be reduced below zero, this fits the data well for most countries over an extended period.

Review questions

EASY

- 1 (a) Define the aggregate demand schedule. (b) How does a fiscal expansion affect the schedule under a flexible inflation target? (c) How would the central bank have to change monetary policy to hit its given inflation target in the long run?
- 2 **Common fallacies** Why are these statements wrong? (a) Fiscal expansion can increase output for ever.
(b) Higher inflation always reduces output.
- 3 An economy has the choice of having half its workers make annual wage agreements every January, and the other half make annual wage agreements every July, or instead forcing everyone to make their annual agreement on 1 July. Which system is likely to induce greater wage flexibility during a period of a few months and during a period of several years?
- 4 How do the following affect the short-run supply schedule, and hence output and inflation in the short run: (a) a higher tax rate; (b) higher labour productivity?
- 5 Which of the following statements is correct? (a) Inflation targeting implies the central bank can ignore what is happening to output. (b) Inflation targeting implies nominal interest rates will typically rise by more than the rise in inflation. (c) Inflation targeting was immediately abandoned once the financial crash of 2009 occurred.

MEDIUM

- 6 Suppose opportunities for investing in high-tech applications boost aggregate demand in the short run but aggregate supply in the long run. Using *AS* and *AD* schedules, show why output might rise *without* much inflation.
- 7 OPEC raises the price of oil for a year but then an increase in the supply of oil from Russia bids oil prices back down again. Contrast the evolution of the economy if monetary policy follows: (a) a fixed interest rate or (b) flexible inflation targeting.
- 8 Distinguish between adjustment in the UK (small open economy, flexible exchange rate) and the US (large economy, international trade a much smaller proportion of its GDP).
- 9 Use the Taylor rule $r - r^* = (1 + a)(\pi - \pi^*) + b(Y - Y^*)$ to answer the following questions: (a) What does the long-run target for the nominal interest rate depend on? (b) In the nominal interest version of the Taylor rule, what happens when there is an increase in inflation? (c) What do the absolute and relative sizes of both the parameters a and b respectively tell us?
- 10 In 2007, the governor of the Bank of England had to write to the Chancellor of the Exchequer to explain why UK inflation had exceeded the target range laid down by the Chancellor. (a) Why were these difficult circumstances? (b) Was the letter proof that the Bank of England was unable to keep inflation in check?
- 11 ‘Central banks, by focusing too much on the inflation rate for goods and services, neglected important signals from asset prices that risk-taking had

become excessive.' Do you agree? What is this likely to imply in future?

HARD

- |2 Imagine that the UK adopts the euro, and interest rates are set by the European Central Bank. (a) Are euro interest rates likely to be adjusted to help stabilize either UK inflation or UK output? (b) What automatic mechanisms, if any, can still achieve these outcomes? (c) Would UK fiscal policy be able to help more?
- |3 Use Figure 21.8 to explore how the collapse of bank lending to companies affects short-run supply curves, and show how the adjustment process subsequently occurs.
- |4 **Essay question** 'Climate change is essentially a permanent adverse supply shock. Production costs will rise; potential output will fall. If the private sector fails to adjust, then either monetary or fiscal policy will have to reduce aggregate demand to the required lower level.' Discuss.

-
- 1 Across countries, higher inflation is often matched by equally higher nominal interest rates, leaving real interest rates roughly constant. This reflects the relative constancy of i^* in the long run. For short-run data for a single country, nominal interest rates vary more than inflation, reflecting the central bank behaviour embodied in Figure 21.1. Recognizing that interest rates must rise sharply when inflation increases has been a key breakthrough of monetary policy design in the last two decades.
 - 2 A similar AD schedule exists if, instead, the central bank pursues a money supply target. For a given path of nominal money M , higher inflation, by raising prices more, reduces the real money supply M/P by more. With lower real money supply, interest rates rise to reduce real money demand and maintain money market equilibrium. Higher real interest rates reduce aggregate demand, just as in Figure 21.2. Under a monetary target, interest rates rise because inflation has reduced the real money supply. Under inflation targeting, interest rates rise in direct response to inflation itself, and the real money supply is then reduced to make this an equilibrium. Either way, higher inflation induces higher real interest rates and lower aggregate demand.
 - 3 A similar analysis applies under monetary targeting. Suppose this is 2 per cent annual growth in nominal money. Long-run inflation will also be 2 per cent. A fall in aggregate demand bids down wage and price growth *below what they would have been*. With inflation below 2 per cent but an unchanged nominal money growth of 2 per cent, the real money supply expands. This causes a fall in real interest rates and boosts aggregate demand back to potential output. Thereafter, money and prices both grow at 2 per cent. The real money supply is permanently higher and real interest rates permanently lower.
 - 4 Looser monetary policy shifts the ii schedule to the right in Figure 21.1. However, once long-run equilibrium is restored, i^* must be unaltered: since aggregate supply is eventually unaltered, aggregate demand cannot eventually change. The only way for the central bank to loosen monetary policy without changing i^* is to accept a higher inflation target π^* .
 - 5 And this finally explains why in Figure 21.1 the central bank does not simply choose a vertical ii schedule at the target inflation rate. When adjustment is sluggish and supply shocks occur, this would imply big swings in output.
 - 6 Paul Krugman's *New York Times* column of 10 October 2009 is reproduced on his blog at <http://krugman.blogs.nytimes.com>.

CHAPTER 22

Inflation, expectations and credibility

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 explain the quantity theory of money
- 2 discuss how nominal interest rates reflect inflation
- 3 analyse seigniorage, the inflation tax and why hyperinflations occur
- 4 assess when budget deficits cause money growth
- 5 explain the Phillips curve
- 6 analyse inflation expectations
- 7 evaluate the costs of inflation
- 8 understand central bank independence and inflation control
- 9 analyse how central banks set interest rates

Sustained inflation is a recent phenomenon. Before 1939, prices rose in some years but fell in others. The UK price level was not substantially higher in 1950 than its level of 1750. Since some years experienced rapid inflation, others must have experienced rapid deflation; that is, a fall in prices. Figure 22.1 confirms this, over an extended period. During this period, the UK was on the gold standard, which anchored money creation to gold reserves. In the long run, money grew slowly, and so did prices. During 1750–1945 falling prices were nearly as likely as rising prices.

Since the Second World War, things have been very different. The world abandoned the gold standard and adopted, first, paper money and, then, electronic money; both could be increased at will by the central bank. There were always economic and political reasons to increase the money supply – and rarely reasons to reduce it. The post-war price level has never fallen: the graph in Figure 22.1 never falls below the horizontal axis after the 1930s. Since 1950 the price level has risen more than 20-fold; that is,

more than its rise over the previous three centuries. This story applies in most advanced economies.

We start with the causes of **inflation**, then examine its effects, which partly depend on whether inflation was anticipated or took people by surprise. We contrast costs that inflation imposes on individuals and costs it imposes on society as a whole. We conclude by considering what the government can do about inflation.

Inflation is a rise in the price level. Pure inflation means that prices of goods and inputs rise at the same rate.

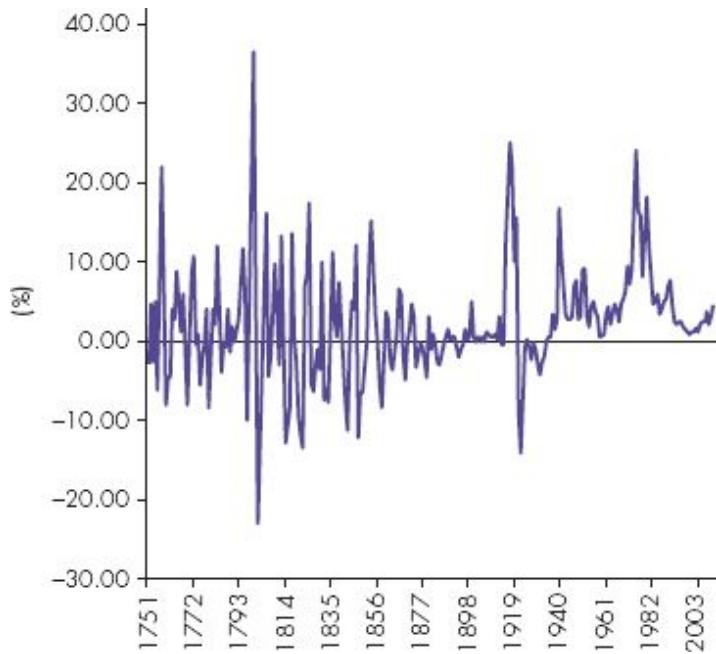


Figure 22.1 The annual UK inflation rate, 1751–2003

Source: www.whatsthecost.com/historic.cpi.aspx.

22.1 Money and inflation

What is the link between nominal money and the price level, and hence between money growth and inflation?

People demand money because of its purchasing power over goods. They demand *real* money. Suppose real income is Y and the interest rate is r . The stock of real money demanded rises if (a) Y increases or (b) r falls.

Conversely, lower income or a higher interest rate leads to a fall in the stock of real money demanded. If L denotes real money demand, we can summarize as

$$M/P = L(Y, r) \quad (1)$$

the left-hand side being real money supplied, the right-hand side real money demanded, depending on income and interest rates. In money market equilibrium, real money supply and demand are equal. Flexible interest rates maintain continuous money market equilibrium. Equation (1) always holds.

CONCEPT 22.1

THE QUANTITY THEORY OF MONEY: $MV = PY$

The velocity of circulation V is nominal income PY divided by nominal money M . If prices adjust to keep real output at potential output Y^* , assumed constant, M and P must move together, *provided velocity V stays constant*. Velocity is the speed at which the stock of money is passed round the economy as people transact. If everyone holds money for less time and passes it on more quickly, the economy needs less money relative to nominal income. How do we assess whether velocity is constant, as the simple quantity theory requires?

The **real money supply** M/P is the nominal money supply M divided by the price level P .

The quantity theory equation implies $M/P = Y/V$. The left -hand side is the **real money supply**. The right-hand side must be real money demand. It rises with real income and falls with velocity. But real money demand rises with real income and falls with nominal interest rates. Hence velocity just measures the effect of interest rates on real money demand. Higher nominal interest rates reduce real money demand. People *hold* less money relative to income. Velocity rises.

While inflation and nominal interest rates are rising, velocity is rising. But if inflation and nominal interest rates settle down at a particular level, velocity is then constant. Thereafter, the simple quantity theory once more applies.

This assumes prices are fully flexible. In the short run, if prices are sluggish, changes in nominal money change the real money supply. Changes in nominal money are not matched by changes in prices. The quantity theory of money will fail in the short run.

If nominal wages and prices adjust slowly in the short run, higher nominal money supply M leads initially to a higher real money stock M/P since prices P have not yet adjusted. The excess supply of real money bids down interest rates. This boosts the demand for goods. Gradually this bids up goods prices. In the labour market, wages start to rise.

After complete adjustment of wages and prices, a one-off rise in nominal money leads to an equivalent one-off rise in nominal wages and prices. Output, employment, interest rates and real money revert to their original levels. After adjustment is complete, the demand for real money is unchanged. Hence the price level changes in proportion to the original change in the nominal money stock.

The **quantity theory of money** says that changes in nominal money lead to equivalent changes in the price level (and money wages), with no effect on output and employment.

The theory is over 500 years old and may date from Confucius. The **quantity theory** is espoused by monetarists, who argue that *most* changes in prices reflect changes in the nominal money supply.

The theory must be interpreted with care. If the demand for real money is constant, the supply of real money must be constant: changes in nominal money are matched by equivalent changes in prices. This raises two issues: (a) even if the demand for real money is constant, do changes in nominal money cause changes in prices or vice versa; and (b) is the demand for real money constant?

Money, prices and causation

Suppose the demand for real money is constant over time. Money market equilibrium implies the real money supply M/P is constant. Monetary policy could fix the nominal money supply M , in which case money M determines prices P to get the required level of M/P implied by money demand.

Conversely, monetary policy may choose a target path for the price level P . Changes in this path then cause changes in the nominal money supply to achieve the required real money supply. Equation (1) says prices and money are correlated, but is agnostic on which causes which. That depends on the form of monetary policy pursued. With an intermediate target for nominal money, the causation flows from money to prices. With a target for prices or inflation, the causation flows the other way.

The leading monetarist Professor Milton Friedman always said that inflation is a monetary phenomenon. Sustained price increases, what we call inflation, are possible only if nominal money is also growing. It is always an option to change monetary policy and stop printing money. Sooner or later prices have to stop rising. Take away the oxygen and the fire goes out.

Is real money demand constant?

Real money demand depends on income Y and the cost of holding money (the spread between interest rates on assets and interest earned, if any, while holding money). Countries with sustained income growth will experience a sustained increase in the demand for real money; countries growing more slowly will have slower growth in real money demand.

Second, countries experiencing different degrees of financial competition will face different costs of holding money. Greater banking competition bid up interest paid on bank deposits, boosting money demand. Changing degrees of banking competition will lead to changes in real money demand.

Third, countries with high inflation are likely to face high nominal interest rates, larger spreads and a higher cost of holding money. Since inflation can become very large, this is potentially the most important reason why real money demand may change. We study this effect in the next section.

To sum up, even once equilibrium is restored, real money demand may be steadily changing, requiring a break in the one-for-one link between money and prices. However, *if* real income and interest rates are unaltered,

changes in nominal money would eventually be accompanied by equivalent changes in nominal wages and prices.

Inflation

So far we have studied levels. Now we focus on rates of change. The growth in real money demand must equal the growth in real money supply; namely, the excess of nominal money growth over the growth in prices. Hence,

$$\text{Nominal money growth} = \text{real money demand growth} + \text{inflation rate}$$

Since real income and interest rates *usually* change only a few percentage points a year, real money demand usually changes slowly.¹ The essential insight of the quantity theory of money is that real variables usually change slowly.

Large changes in one nominal variable (money) are then accompanied by large changes in other nominal variables (prices, nominal wages) to keep real money (and real wages) at their equilibrium values. This is a useful first look at inflation, but we simplified too much.

22.2

Inflation and interest rates

The **Fisher hypothesis** says higher inflation leads to similarly higher nominal interest rates.

Figure 22.2 shows interest and inflation rates for 15 countries in 2013. Countries with high inflation have high interest rates. An extra percentage point of inflation is accompanied on average by a nominal interest rate nearly one percentage point higher, a proposition first suggested by Professor Irving Fisher. By definition,

$$\text{Real interest rate} = [\text{nominal interest rate}] - [\text{inflation rate}]$$

The **Fisher hypothesis** says that *real* interest rates do not change much. If they did, there would be large excess supply or demand for loans. Hence, higher inflation is largely offset by higher nominal interest rates to stop the real interest rate changing much. Figure 22.2 shows this is a good rule of thumb in reality. Countries, such as Switzerland, with inflation rates close

to zero have correspondingly low nominal interest rates. Turkey and Hungary have interest rates around 5 per cent, largely because their inflation rates are at similar levels. In Pakistan, the levels of both are around 9 per cent, and in Venezuela nominal interest rates are nearly 15 per cent, but then inflation is over 20 per cent.

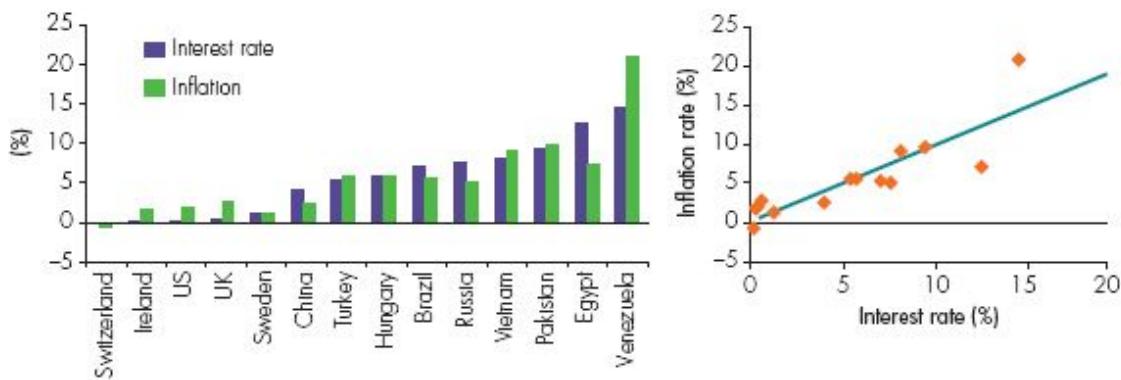


Figure 22.2 Inflation and interest rates, January 2013 (% per annum)

Source: <http://www.economist.com/markets-data>.

Faster nominal money growth leads both to higher inflation and higher nominal interest rates. Hence a rise in the rate of money growth leads to a rise in nominal interest rates. This reduces the demand for real money, requiring money and prices to grow at *different* rates until the real money supply adjusts to the change in real money demand. To show how this works, we study a spectacular example – the German hyperinflation.

Hyperinflation

Hyperinflation is a period of very high inflation.

Bolivian annual inflation reached 11 000 per cent in 1985, Ukraine's inflation topped 10000 per cent in 1993, and in 2009 inflation in Zimbabwe may have exceeded a trillion per cent – its central bank had to resort to printing banknotes denominated in 100 trillion Zimbabwean dollars that were still worth only a few US dollars or UK pounds.

The most famous, and most studied, hyperinflation is that in Germany during 1922–23. Germany lost the First World War. The German government had a big deficit, financed by printing money. Table 22.1 shows what happened. The government had to buy faster printing presses. In the later stages of the hyperinflation, they took in old notes, stamped on another zero, and reissued them as larger-denomination notes in the morning.

	Money	Prices	Real money	Inflation % monthly
January 1922	1	1	1.00	5
January 1923	16	75	0.21	189
July 1923	354	2021	0.18	386
September 1923	227 777	645 946	0.35	2532
October 1923	20 201 256	191 891 890	0.11	29 720

Source: adapted from C. L. Holtfrerich, *Die Deutsche Inflation 1914–23* (Walter de Gruyter, 1980).

If inflation is π and the nominal interest rate is r , the real interest rate is $(r - \pi)$ but the real return on non-interest-bearing cash is $2p$, which shows how quickly the real value of cash is being eroded by inflation. The extra real return on holding interest-bearing assets rather than cash is $(r - \pi) - (-\pi) = r$. The *nominal* interest rate measures the *real* cost of holding cash. Nominal interest rates rise with inflation. In the German hyperinflation the cost of holding cash became enormous.

Table 22.1 shows that German prices rose 75-fold in 1922, and more in 1923. People carried money in wheelbarrows to go shopping. By October 1923, real money holdings were only 11 per cent of their level in January 1922. How did people get by with such small holdings of real cash?

People, paid twice a day, shopped in their lunch hour before the real value of their cash depreciated too much. Any cash not immediately spent was quickly deposited in a bank where it could earn interest. People spent a lot of time at the bank.

The **flight from cash** is the collapse in the demand for real cash when high inflation and high nominal interest rates make it very expensive to hold cash.

What lessons can we draw? First, *rising* inflation and *rising* interest rates significantly reduce the demand for *real* cash. Hyperinflations are a rare example in which a real quantity (real cash) changes quickly and by a lot. Second, and as a result, money and prices can get quite out of line when inflation and nominal interest rates are rising. Table 22.1 shows that prices rose by six times as much as nominal money between January 1922 and July 1923, reducing the real money supply by 82 per cent, in line with the fall in real money demand – a **flight from cash**.

22.3

Inflation, money and deficits

Persistent inflation must be accompanied by continuing nominal money growth. Printing money to finance a large deficit is a source of inflation. Budget deficits may explain why governments have to print money rapidly. Tight *fiscal* policy is needed to fight inflation.

If government debt is low relative to GDP, the government can finance deficits by borrowing. It has enough tax revenue with which to pay interest and repay the debt. For governments with low debt, there may be no relation between their budget deficit and how much money they print. Sometimes they print money; sometimes they issue bonds. We do not expect a close relationship between deficits and money creation in most advanced countries.

Many years of deficits may make government debt large relative to GDP. The government then cannot finance deficits by more borrowing. It has to tighten fiscal policy to shrink the deficit or print money to finance the continuing deficit.

To ensure that the European Central Bank did not face fiscal pressure to print too much money and thus create inflation, members of the Eurozone had to obey the Stability and Growth Pact, which restricts their budget deficits to less than 3 per cent of GDP, except in severe recession. Of course, when severe recession arrived in 2009, budget deficits escalated to 10 per cent and beyond. We never know how binding a commitment will be until a crisis occurs.

Deficits, money growth and real revenue

A hyperinflation is a situation in which fiscal policy is out of control. A government with a persistently high deficit, financed by borrowing, now

has so much debt that nobody will lend it any more. Instead, it prints money to finance its deficit.

Seigniorage is non-inflationary real revenue acquired by the government through its ability to print money.

How much real revenue can the government get by printing banknotes? The government has a monopoly on cash. As a token money, its production cost is tiny relative to its value as money. The government prints money for nothing, then uses it to pay nurses and build roads. Real money demand M/P rises with real income. Long-run growth of real income allows the government some scope to raise M without adding to P . This extra money can be used to finance government spending, and so is as good as tax revenue, without causing inflation. It is called **seigniorage**, and arises from the government monopoly on creating money with a value vastly in excess of its production cost.

A second potential source of real revenue is the inflation tax. Suppose real income and output are constant but that a weak government cannot shrink its budget deficit and now has debt so large that nobody will lend to it. It prints money to cover the budget deficit. If ΔM is the amount of new cash created, this finances an amount of real spending $(\Delta M)/P$, which is the same as $(\Delta M/M) \times (M/P)$, the growth rate of cash multiplied by the real demand for cash. The rise in nominal money must feed into prices sooner or later. Suppose the rate of nominal money growth $(\Delta M/M)$ equals the inflation rate π . Thus,

$$\text{Real revenue from inflation} = \pi \times M/P$$

The **inflation tax** is the effect of inflation in raising real revenue by reducing the real value of the government's nominal debt.

Inflation helps the government by reducing the real value of the non-interest-bearing part of the government debt, namely cash. Think of inflation as the tax rate and real cash as the tax base for the **inflation tax**.

Now for the part that may be new to you. If money growth and inflation rise, does the government get more *real* revenue from the inflation tax? Higher inflation raises nominal interest rates and hence reduces the real demand for cash.

Figure 22.3 shows the answer. At low inflation, real cash demand is high, but the multiple of inflation and real cash demand is small. Similarly, at high inflation, although the inflation tax rate is high, the tax base – real cash demand – is now tiny because nominal interest rates are so high. The multiple of inflation and real cash is again low. Real revenue raised through the inflation tax cannot be increased indefinitely. After a certain point, faster money growth and higher inflation shrink the tax base more than they raise the tax rate.

The figure has two implications. First, if the government needs to cover a particular *real* deficit d by printing money, there may be two rates of money growth and inflation that do the job. Either is a long-run equilibrium in which inflation is constant.

Second, if for political reasons the government has a real deficit as large as D , printing money cannot do the job. The economy explodes into hyperinflation. At high inflation, real cash demand is already low. Raising inflation further causes such a large percentage fall in the tiny demand for real cash that inflation tax revenue falls, the government prints even more cash and the problem gets even worse.

That is how hyperinflation starts. The only solution is to cut the size of the deficit. Often the government does this by defaulting on its debt, which slashes the burden of interest payments.

Notice that the inflation tax applies to cash which has no nominal interest rate to increase in line with inflation. For interest-bearing money, in principle interest rates can rise to protect money holders. Then it is only unforeseen inflation, not incorporated in interest rates, that acts as a tax.

This is one reason we tend to see hyperinflation in more primitive economies, in which cash is very important. In modern European economies, cash is much less important and the potential tax base for the inflation tax is a lot lower.

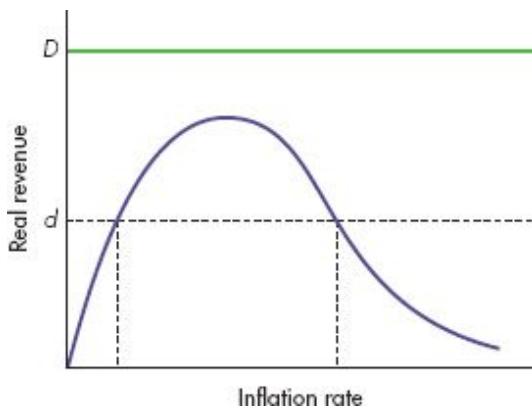


Figure 22.3 Real revenue from the inflation tax

22.4

Inflation, unemployment and output

One of the most famous relationships in post-war macroeconomics is the Phillips curve.

The Phillips curve

In 1958 Professor Phillips of the London School of Economics found a strong statistical relationship between annual inflation and annual unemployment in the UK. Similar relationships were found in other countries. The Phillips curve is shown in Figure 22.4.

The **Phillips curve** shows that a higher inflation rate is accompanied by a lower unemployment rate. It suggests we can trade off more inflation for less unemployment or vice versa.

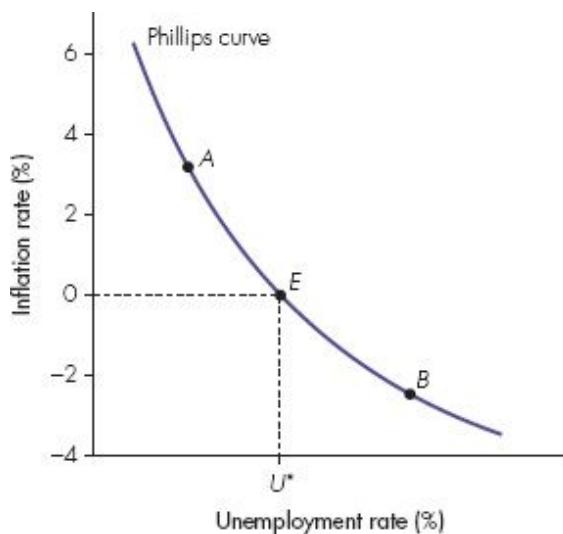
The **Phillips curve** seemed a useful compass for choosing macroeconomic policy. By its choice of fiscal and monetary policy, the government set aggregate demand and hence unemployment. The Phillips curve showed how much inflation then ensued. Higher aggregate demand bid up wages and prices, causing higher inflation but lower unemployment.

The Phillips curve shows the trade-off that people believed they faced in the 1960s. In those days UK unemployment was rarely over 2 per cent of the labour force. But people believed that, if they did the unthinkable and

reduced aggregate demand until unemployment rose to 2.5 per cent, inflation would fall to zero.

Since then there have been years when *both* inflation and unemployment were over 10 per cent. Something happened to the Phillips curve. The next two chapters explain why the simple Phillips curve of Figure 22.4 ceased to fit the facts.

Equilibrium unemployment is not zero, for reasons that we explore in Chapter 23. Suppose equilibrium employment and potential output are fixed in the long run, but there is sluggish wage and price adjustment. Chapter 21 discussed the vertical long-run aggregate supply curve and sloping short-run supply curve, relating output and the price level. These ideas are easily translated from inflation and output to inflation and unemployment.



The Phillips curve shows the trade-off between higher inflation and lower unemployment. In the 1960s people believed that an unemployment rate U^* of 2.5 per cent would be accompanied by zero inflation.

Figure 22.4 The Phillips curve

The vertical long-run Phillips curve

The **natural level of output** and the **natural rate of unemployment** are the long-run equilibrium levels of output and unemployment levels, respectively.

In long-run equilibrium, the economy is at both potential output and equilibrium unemployment. Sometimes these are referred to as the **natural level of output** and the **natural rate of unemployment**.

Both are determined by real things, not nominal things. They depend on the supply of inputs, the level of technology, the level of tax rates, and so on. They do not depend on inflation, provided all prices P and nominal wages W are rising together. Equilibrium unemployment depends on the real wage W/P , as we discuss in Chapter 23.

Just as long-run aggregate supply is vertical at potential output – output is unaffected by inflation – so the long-run Phillips curve is vertical at equilibrium unemployment. Equilibrium unemployment is independent of inflation. Plotting inflation and unemployment, Figure 22.5 shows the long-run Phillips curve vertical at equilibrium unemployment U^* .

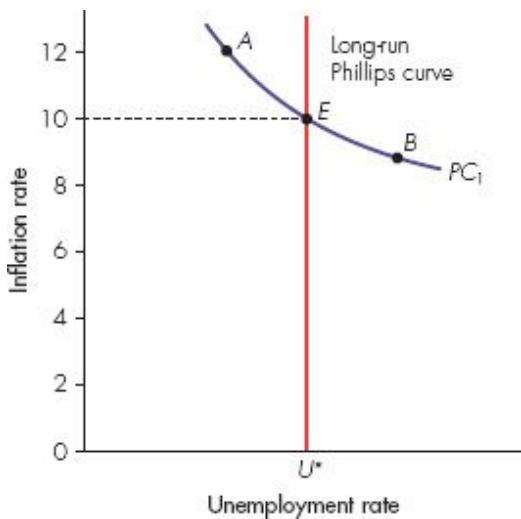
In long-run equilibrium, inflation is constant. People correctly anticipate inflation, and adjust the growth of nominal wages to keep real wages constant, at the real wage required for long-run equilibrium. Similarly, nominal interest rates are sufficiently high to offset inflation and maintain real interest rates at their equilibrium level. Everyone adjusts to inflation because it can be completely foreseen.

Suppose inflation is 10 per cent a year. This is consistent with many forms of monetary policy. We can think of monetary policy as having either a target of 10 per cent annual money growth, or an inflation target of 10 per cent a year, or as a Taylor rule in which the inflation part aims for 10 per cent annual inflation. In Figure 22.5 long-run equilibrium is at E . Inflation is 10 per cent, as everybody expects. Nominal money grows at 10 per cent a year. Unemployment is at its natural rate.

The short-run Phillips curve

Beginning from E , suppose something raises aggregate demand. Unemployment falls, inflation rises and the economy is at A . Then the central bank raises interest rates to achieve its targets (in whichever form), and the economy slowly moves back down the short-run Phillips curve PC_1 from A back to E again. Since interest rates take time to affect aggregate demand, this may take one or two years.

Conversely, beginning from E a downward demand shock takes the economy to B in the short run. The central bank alters interest rates to bring the economy steadily back from B to E .



Since people care about real variables not nominal variables, when full adjustment has been completed people will arrange for all nominal variables to keep up with inflation. The vertical long-run Phillips curve shows that eventually the economy gets back to the natural rate of unemployment U^* , whatever the long-run inflation rate. There is no long-run trade-off between inflation and unemployment. The short-run Phillips curve PC_1 shows short-run adjustment as before. The height of the short-run Phillips curve depends on the rate of inflation and nominal money growth in long-run equilibrium, as shown by the position of point E on the long-run Phillips curve.

Figure 22.5 The long-run Phillips curve

The **short-run Phillips curve** corresponds to the short-run supply curve for output. Given inherited wages, higher prices make firms supply more output and demand more workers. For any level of last period's prices, higher prices today imply higher inflation today. In Chapter 21, the height of the short-run aggregate supply curve depended on the inherited growth rate of nominal wages. Similarly, the height of the short-run Phillips curve reflects inherited nominal wage growth.

The **short-run Phillips curve** shows that, in the short run, higher unemployment is associated with lower inflation. The height of the short-run Phillips curve reflects expected inflation. In long-run equilibrium at E, expectations are fulfilled.

When workers and firms expect high inflation, they agree a large rise in nominal wages. If inflation turns out as expected, real wages are as forecast and the nominal wage growth was justified. If inflation is higher than expected, real wages are lower than planned. Firms supply more

output and demand more labour. High inflation (relative to expectations) goes with lower unemployment. The short-run Phillips curve slopes down. Its height reflects the inflation expectations embodied in the inherited wage agreement.

This explains why most economies had high inflation at each unemployment rate in the 1970s and 1980s: the short-run Phillips curve had shifted upwards. Governments were printing money at a faster rate than before. The long-run equilibrium inflation rate was high, and expected to be so.

The point E lay further up the long-run Phillips curve in Figure 22.5. The short-run Phillips curve through this point was much higher than the short-run Phillips curve in the data originally studied by Professor Phillips. The 1970s and 1980s were periods of high inflation. The original Phillips curve data had been for a period of much lower inflation.

Figure 22.6 confirms the complete correspondence between the aggregate supply schedules of the previous chapter and the Phillips curves in this chapter. In the long run, the economy is at potential output Y^* and equilibrium unemployment U^* . Suppose the inherited level of inflation expectations is the level p . The economy then faces the short-run aggregate supply schedule SAS in the left-hand panel and the short-run Phillips curve PC in the right-hand panel – one implies the other. Initially the economy is at point A in both panels.

A fall in aggregate demand shifts the economy to point A' in both panels – there is less demand, inflation falls, output falls and unemployment rises. Conversely, the short-run effect of a rise in aggregate demand takes the economy to point A'' in both panels.

If inflation expectations are firmly anchored, eventually the economy will return to point A in both panels. However, if having first experienced an increase in aggregate demand and a move to point A'' , inflation expectations then increase permanently from π to π_1 , the short-run aggregate supply curve shifts up from SAS to SAS_1 and the short-run Phillips curve from PC to PC_1 .

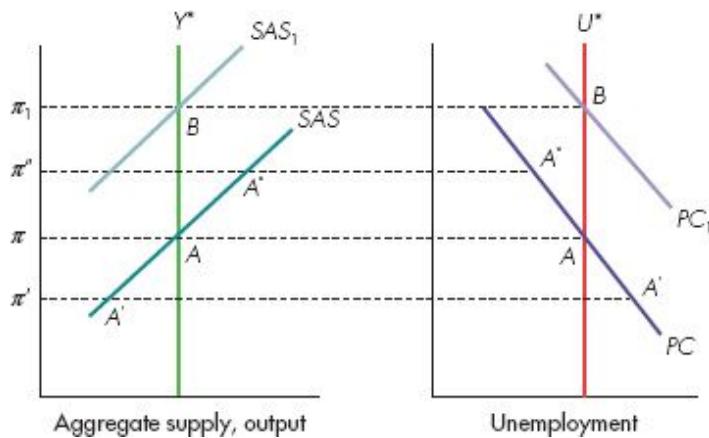


Figure 22.6 The correspondence of aggregate supply and the Phillips curve

We could also use Figure 22.6 to explore permanent supply shocks. For example, an increase in work incentives would shift the vertical long-run Phillips curve to the left, reducing equilibrium unemployment. Correspondingly, long-run aggregate supply of output would increase, shifting the vertical long-run aggregate supply curve to the right.

We draw two conclusions. First, it was wrong to interpret the original Phillips curve as a *permanent* trade-off between inflation and unemployment. It was the temporary trade-off, corresponding to a particular short-run aggregate supply schedule, while the economy adjusted to a demand shock.

Second, the speed with which the economy moves back along the Phillips curve depends on two things: the degree of flexibility of nominal wages and hence prices; and the extent to which monetary policy adjusts interest rates to restore demand more quickly. Complete wage flexibility would restore the vertical Phillips curve and the vertical aggregate supply curve. Rapid adjustment of interest rates would offset the demand shock, restoring output, unemployment and inflation to their long-run equilibrium levels.

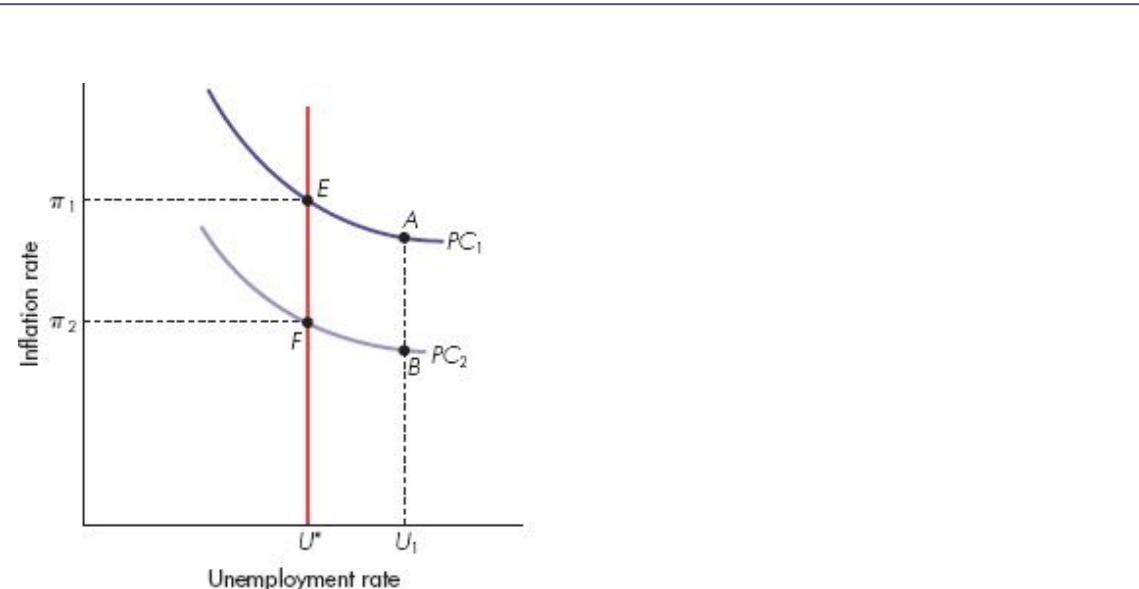
Extreme monetarists believe that wage flexibility is very high. In the extreme version, it is only the fact that workers make annual wage settlements that prevents the economy always being in long-run equilibrium. Changes in aggregate demand unforeseen when nominal wages were set mean that wages and prices are temporarily at the wrong level. But such mistakes are rectified as soon as wages are renegotiated.

If wage and price adjustment are more sluggish than this, full employment is not immediately restored. However, we know from the previous chapter that monetary policy can completely compensate for a demand shock once it has been diagnosed. Nor is there any conflict between stabilizing inflation and stabilizing output or employment. Such conflicts arise only in response to supply shocks.

We have made considerable progress in understanding the Phillips curve, but there is more still to study. First, we need to analyse changes in long-run inflation expectations, which shift the short-run Phillips curve. Second, we need to examine supply shocks. Temporary supply shocks also shift the short-run Phillips curve. Permanent supply shocks alter equilibrium unemployment and shift the long-run Phillips curve.

Expectations and credibility

Figure 22.7 puts this apparatus to work to discuss what happens when a new government is elected with a commitment to reduce inflation. The economy begins in long-run equilibrium at E , facing the short-run Phillips curve PC_1 . Nominal money, prices and money wages are all rising at the rate π_1 .



Beginning at E , the target inflation rate is cut from π_1 to π_2 . Having expected inflation π_1 , nominal wage growth has been too high. Firms cut back output and employment and the economy moves to A . If the new policy is credible, the next wage settlement reflects lower inflation expectations, the short-run Phillips curve shifts to PC_2 and the economy moves from A to B . Thereafter, it slowly adjusts along PC_2 to F . However, if people doubt that the new

tough policy will be sustained, nominal wages may keep growing at π_1 . The short-run Phillips curve remains PC_1 . Unemployment stays high, and inflation refuses to fall.

Figure 22.7 Expectations and credibility

The government wants to reduce inflation to p_2 to reach point F . The day the government is elected it announces a cut in the inflation target from π_1 to π_2 .

Overnight, firms inherit nominal wage increases that had anticipated the old inflation rate π_1 . They have little scope to reduce inflation. If inflation does fall, real wages are now too high. Firms reduce output and employment. Inflation falls a little and unemployment rises. The economy moves along the short-run Phillips curve PC_1 to A .

What happens next? In the good scenario, workers believe the tighter monetary policy will last. The next wage bargain is based on inflation expectations π_2 . The short-run Phillips curve shifts down to PC_2 and the economy moves from A to B . Inflation falls quickly. The economy then moves slowly along PC_2 from B to F .

Now for the bad scenario. When the economy first reaches A , workers do not believe that the tough new monetary policy will last. They think p_1 will remain the inflation rate in the long run. Thinking inflation will remain high, workers do not reduce nominal wage growth. They believe PC_1 not PC_2 will be relevant. A self-fulfilling prophecy is an expectation that creates the incentive to make it come true.

Suppose workers are wrong. Although nominal wages grow at π_1 , the tough policy lasts and actual inflation is below π_1 . Real wages rise and unemployment gets worse without much fall in inflation. The worse the slump becomes, the more likely is the government to give in, easing monetary policy to boost aggregate demand again. A belief that the government's nerve will crack can become a **self-fulfilling prophecy**.

A **self-fulfilling prophecy** is an expectation that creates the incentive to make it come true.

The economy stays on PC_1 and the attempt to reduce inflation fails. Gradually the economy moves back along PC_1 to equilibrium at E .

This explains why governments go to such lengths to commit to tight monetary policy. The sooner people accept that long-run inflation will be low, the sooner nominal wage growth will slow. Making central banks independent is an institutional reform designed to increase the credibility of monetary policy by insulating it from short-term political expediency.

Figure 22.7 was used to describe a fall in inflation expectations, but the opposite is also possible. By 2010 many people were wondering whether Western governments would indeed curtail government spending and raise taxes in order to bring budget deficits under control. If future governments are weak, they may resort to money creation instead. We could then envisage a shift upwards in the short-run Phillips curve in Figure 22.7 as inflation expectations increase.

In assessing how likely this is, people will look at the independence of the central banks and whether they are prepared to use this to adhere to the monetary targets they have been set. In principle, central banks should be prepared to raise interest rates if they foresee any systematic increase in inflation above its target level.

CONCEPT 22.2

EXPECTATIONS FORMATION AND THE ACCELERATIONISTHYPOTHESIS

How should economists model inflation expectations? One solution is simply to treat inflation expectations as exogenous or autonomous – given from outside the model. We analyse conditional on a particular assumption, then explore the consequences of changing that assumption. This is simple, but not very satisfactory – if the Phillips curve is telling us about inflation and unemployment, should not the people who populate the model be reacting to the same information when they form expectations?

For many years, economists relied on the assumption of adaptive (or extrapolative) expectations, which assumes that people use the past history of inflation to forecast its future values. If at any time their forecast turns out to be too low (high), they revise upwards (downwards) their inflation expectations for the following period. Eventually, in long-run equilibrium, expectations and actuality have converged.

This expectations assumption gave rise to a clear prediction. If governments could *accelerate* the rate of money creation, steadily increasing its rate, they could manufacture inflation that kept ahead of people's expectations for a *sustained* period. Workers would repeatedly settle for nominal wage growth that did not keep place with inflation, reducing the real wage, and promoting higher employment and output. The figure below illustrates.

In the long run, the Phillips curve is vertical at the equilibrium unemployment rate U^* . Eventually, the labour market can adjust to any level of sustained inflation.

Initially, suppose expected inflation is zero. The short-run Phillips curve $SRPC_1$ would be compatible with unemployment at U^* if actual inflation turned out to equal the expected inflation rate of zero. However, the government can artificially lower actual unemployment to U_1 by generating inflation p_1 . Since this exceeds the zero inflation expectations embodied in wage settlements, it temporarily reduces the value of real wages until the labour market has time to adjust.

Next period, everybody has raised their inflation expectations a bit, and nominal wages increase a bit. The short-run Phillips curve $SRPC_2$ captures the higher inflation expectations. Its intersection with the long-run Phillips curve shows how much inflation is now being expected. However, by boosting the economy even more strongly, the government can achieve an actual inflation rate p_2 that again exceeds expectations and lowers real wages. Again, actual unemployment is reduced to U_1 . The next wage settlement recognizes that expectations have twice been too low. The short-run Phillips curve reflects another increase in inflation expectations. A government determined to keep unemployment down could still achieve actual inflation π_3 , and again engineer a reduction in real wages that reduces unemployment to U_1 .

The conclusion is that, if expectations adjust adaptively in this way, a determined government could reduce unemployment for a sustained period by a sustained sequence of inflation surprises. But as inflation expectations rise, it takes ever higher inflation to keep surprising the labour market.

Conversely, if a new inflation-hating government takes over, it can gradually bash down inflation expectations, and gradually shift the short-run Phillips curve down again. Each time, it will generate less

inflation than had been expected, inducing a subsequent downward revision in expectations, till eventually $SPRC_1$ can again be reached. During the transition, each episode entails *less* inflation than expected, and hence implies real wages that are higher than intended when nominal wage settlements were reached. Hence, there is a period of abnormally high unemployment while the $SRPC$ is gradually being shifted downwards.

The **accelerationist hypothesis** states that, by accelerating inflation faster than workers' expectations can adjust, the government can depress real wages and achieve abnormally low unemployment for a sustained period.

Adaptive expectations are a model of limited rationality. People get fooled in the short run, but eventually get things right once long-run equilibrium is reached. Some economists think people are smarter than this. The **accelerationist hypothesis** relies on fooling people period after period with the same trick. If people grasp what the government is doing, they will anticipate the additional inflation and the power to surprise will be weakened or eliminated.

Supply shocks

In the long run, other things equal, the Phillips curve is vertical at equilibrium unemployment U^* . But other things are unequal, and U^* is not constant. In terms of Figure 22.6, a rise in equilibrium unemployment shifts the vertical long-run Phillips curve to the right. Changes in equilibrium unemployment reflect **permanent supply shocks**.

A **permanent supply shock** affects equilibrium unemployment and potential output.

The short-run Phillips curve can shift for two reasons. Inherited nominal wage growth changes if inflation expectations change, as analysed in Figure 22.7. Alternatively, a change in firms' desired supply of output and demand for workers, for a given rate of inherited nominal wage growth, shifts the short-run Phillips curve. Examples include a change in oil prices, regulations or tax rates.

SHORT-RUN AND LONG-RUN PHILLIPS CURVES

Consider the short-run Phillips curve:

$$\pi = \pi^e - b(U - U^*) \quad b > 0 \quad (1)$$

where U and U^* are, respectively, actual unemployment and equilibrium unemployment, π is inflation and π^e is expected inflation. When actual and equilibrium unemployment coincide, inflation is determined by the inherited level of inflation expectations, which therefore determines the height of the short-run Phillips curve. When expected inflation is higher, the entire short-run Phillips curve is higher. The parameter b determines the slope of the short-run Phillips curve. The larger is b , the steeper is the short-run curve. Because b is a constant, in this example $SRPC$ has a constant negative slope. In the long run, any level of inflation expectations can prevail when U and U^* coincide. All we know is that then actual and expected inflation coincide along the long-run Phillips curve. Any actual inflation rate is compatible eventually with U^* . $LRPC$ is vertical.

Suppose people believe the central bank will try to stabilize inflation at 2 per cent, but that this cannot be achieved overnight. People therefore expect:

$$\pi^e = 0.02 - a(U - U^*) \quad a > 0 \quad (2)$$

When unemployment exceeds U^* , people expect inflation to be lower than normal; when unemployment is below U^* , people expect a boom to be leading to above normal inflation.

Combining equations (1) and (2):

$$\pi = 0.02 - (a + b)(U - U^*) \quad (3)$$

In this example, when unemployment has reverted to its long-run equilibrium, inflation is then 2 per cent. In the short run, inflation is affected by deviations of unemployment from equilibrium unemployment not only because this affects wages, prices and inflation, but also because it has a second effect on inflation expectations themselves. Exactly how inflation expectations adjust over time in the real world is a subject of continuing controversy.

We can also examine the relationship between the Phillips curve in this chapter and our discussion of aggregate supply in the previous chapter. If Y^* is the level of potential output, determined in the long run by aggregate supply, and the short-run output gap is $(Y - Y^*)$, then demand-driven cyclical fluctuations in output should lead to corresponding cycles in employment, and hence in unemployment:

$$Y - Y^* = -h(U - U^*) \quad h > 0 \quad (4)$$

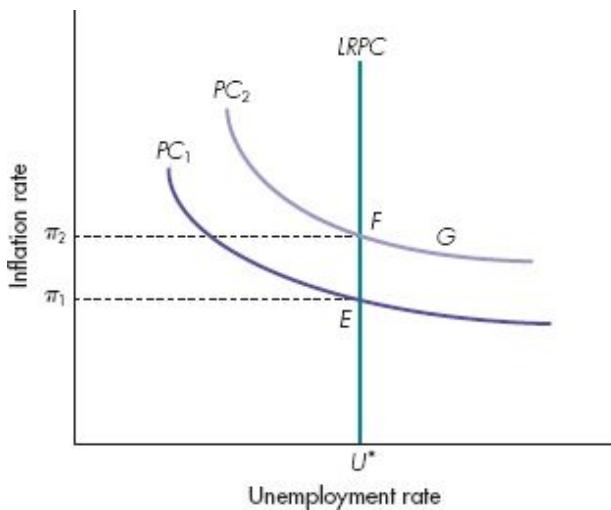
When output exceeds potential output, unemployment lies below its natural rate. Combining equations (1) and (4):

$$Y = Y^* + (h/b)(\pi - \pi^e) \quad (5)$$

which we can interpret as the short-run aggregate supply curve of the previous chapter. For a given level of inherited inflation expectations, higher current inflation induces firms to supply additional current output. In the long run, inflation and expected inflation coincide, and current output simply equals potential output.

Figure 22.8 shows an adverse **temporary supply shock**. A temporary supply shock leaves these long-run values unaffected, but shifts the short-run Phillips curve and the short-run aggregate supply schedule for output. The short-run Phillips curve shifts up, from PC_1 to PC_2 . If monetary policy accommodates the shock, the target inflation rate rises from π_1 to π_2 . The economy moves from E to F with no change in output or unemployment, but at the cost of higher inflation. Eventually the shock wears off, since it is temporary, and the economy reverts to E , with another accommodating change in monetary policy.

A **temporary supply shock** leaves these long-run values unaffected, but shifts the short-run Phillips curve and the short-run aggregate supply schedule for output.



An adverse but temporary supply shock shifts PC_1 to PC_2 without affecting LRPC.

Beginning from E monetary policy can accommodate the shock, moving to F. If interest rates are raised to prevent inflation rising as high as π_2 , the fall in demand raises unemployment. At G the economy experiences stagflation, both high inflation and high unemployment.

Figure 22.8 Temporary supply shocks

Alternatively, monetary policy may *not* fully accommodate the supply shock. In Chapter 21, we showed that this would mean higher inflation *and* lower output. Now, the analogue is higher inflation *and* higher unemployment – **stagflation**. To prevent inflation shifting up by as much as the vertical shift up in the short-run Phillips curve, monetary policy makes sure that aggregate demand falls a bit. Hence inflation rises a bit and unemployment rises a bit. The economy moves from E to G in Figure 22.7. Output stagnates despite higher inflation.

Stagflation is high inflation and high unemployment, caused by an adverse supply shock.

Again, the credibility of policy is crucial. If workers think the government, frightened of high unemployment, will accommodate any shock, large wage rises buy temporarily higher real wages until prices adjust fully. And in the long run, monetary policy is loosened to maintain aggregate demand at full employment, so there is little danger of extra unemployment.

Once a government proves that it will not accommodate shocks, nominal wage growth slows. Workers then fear that higher wages will reduce demand and price workers out of a job.

22.5 The costs of inflation

People dislike inflation, but why is it so bad? Some reasons commonly given are spurious.

Inflation illusion?

People have **inflation illusion** if they confuse nominal and real changes. People's welfare depends on real variables, not nominal variables.

It is wrong to say that inflation is bad because it makes goods more expensive. If *all* nominal variables rise at the same rate, people have larger nominal incomes and can buy the same physical quantity of goods as before. If people realize that prices have risen but forget that nominal incomes have also risen, they have **inflation illusion**. It is real incomes that tell us how many goods people can afford to buy.

A second mistake is more subtle. Suppose there is a sharp rise in the real price of oil. Oil-importing countries are worse off. Domestic consumption per person has to fall. It can fall in one of two ways.

If workers do not ask for ‘cost-of-living’ wage increases to cover the higher cost of oil-related products, real wages fall. Nominal wages buy fewer goods. Suppose too that domestic firms absorb higher oil-related fuel costs, not passing on these costs in higher prices. There is no rise in domestic prices or nominal wages. The domestic economy has adjusted to the adverse supply shock without inflation. People are worse off.

Suppose instead that people try to maintain their old standard of living. Workers claim cost-of-living rises to restore their real wages, and firms protect their profit margins by raising prices in line with higher wage and fuel costs. There is a lot of domestic inflation, which the government accommodates by printing extra money. Eventually the economy settles down in its new long-run equilibrium position.

People must still be worse off. The rise in the real oil price has not disappeared. It still takes more domestic exports, using resources no longer available to make output for domestic consumption, to pay for the more expensive oil imports. In the new long-run equilibrium, wages do not quite keep up with higher prices, and prices do not quite keep up with higher

costs. The market brings about the required fall in real domestic spending, letting resources go into exports to pay for the more expensive oil imports.

People notice (a) rising prices and (b) lower real incomes, but draw the wrong conclusion. Inflation did not make them worse off, higher oil prices did. Inflation is a symptom of the initial refusal to accept the new reality.

We now turn to better arguments about the cost of inflation. Our discussion has two themes. First, was the inflation fully expected, or were people surprised? Second, do our institutions, including regulations and the tax system, let people adjust fully to inflation once they expect it? The cost of inflation depends on the answer to these two questions.

Complete adaptation and full anticipation

Imagine an economy with annual inflation of 10 per cent for ever. Everybody anticipates it. Nominal wages grow and nominal interest rates incorporate it. Real wages and real interest rates are unaffected. The economy is at full employment. Government policy is fully adjusted. Nominal taxes are changed every year to keep real tax revenue constant. Nominal government spending rises at 10 per cent a year to keep real government spending constant. Share prices rise with inflation to maintain the real value of firms. The tax treatment of interest earnings and capital gains is adjusted to reflect inflation. Pensions and other transfer payments rise every year, in line with expected inflation.

This economy has no inflation illusion. Everyone has adjusted to it. This explains the long-run vertical Phillips curve in the previous section. Is complete adjustment possible?

Shoe-leather costs of inflation are the extra time and effort in transacting when we economize on holding real money.

Menu costs of inflation are the physical resources needed for adjustments to keep real things constant when inflation occurs.

Nominal interest rates usually rise with inflation to preserve the real rate of interest. But the nominal interest rate is the opportunity cost of holding cash. When inflation is higher, people hold less real cash. Society uses money to economize on the time and effort involved in undertaking transactions. High nominal interest rates make people economize on real money – thus incurring **shoe-leather costs**. Using more resources to

transact, we have fewer resources for production and consumption of goods and services.

When prices rise, price labels have to be changed. Menus are reprinted to show the higher price of meals. The faster the rate of price change, the more often menus must be reprinted if real prices are to remain constant. Among the **menu costs of inflation** is the effort of doing mental arithmetic. If inflation is zero, it is easy to see that a beer costs the same as it did three months ago. When inflation is 25 per cent a year, it takes more effort to compare the real price of beer today with that of three months ago. People try to think in real terms, but the mental arithmetic involves time and effort.

How big are menu costs? Supermarkets easily change price tags. The cost of changing parking meters, pay telephones and slot machines is larger. In countries with high inflation, pay phones usually take tokens whose price is easily changed without having physically to alter the machines.

Even when inflation is perfectly anticipated and the economy has fully adjusted to it, we cannot avoid shoe-leather and menu costs. These costs are big when inflation is high, but may not be too big when inflation is moderate. However, if we cannot adjust to expected inflation, the costs are then larger.

Fully anticipated inflation when institutions do not adapt

Assume inflation is fully anticipated but institutions prevent people fully adjusting to expected inflation. Inflation now has extra costs.

Fiscal drag is the rise in real tax revenue when inflation raises nominal incomes, pushing people into higher tax brackets in a progressive income tax system.

Taxes

Tax rates may not be fully inflation-adjusted. One problem is **fiscal drag**. Suppose income below £10 000 was untaxed, but you pay income tax at 50 per cent on all income over that amount. Initially, you earn £15 000 and pay income tax of £2500. After ten years of inflation, all wages and prices double but tax brackets and tax rates remain as before. You now earn £30

000. Paying tax at 50 per cent on the £20 000 of taxable income, you pay tax of £10 000. Wages and prices only doubled. Your nominal tax rose from £2500 to £10 000. Fiscal drag raised the real tax burden. The government gained from inflation. You lost.

To be inflation-neutral, nominal tax brackets must rise with inflation. When prices double, the nominal tax exemption must rise, from £10 000 to £20 000. At your new nominal income of £30 000, you pay £5000 tax, exactly twice in nominal terms what you paid before prices increased. In real terms, nothing has changed.

Percentage taxes on value, such as VAT, automatically raise nominal tax revenue in line with the price level. However, *specific* duties, such as £5 on a bottle of whisky, must be raised as the price level rises.

Taxing capital

Income tax on interest income is also affected by inflation. Suppose there is no inflation. Nominal and real interest rates are both 4 per cent. With a 40 per cent tax rate, the after-tax real return on lending is 2.6 per cent.

Now suppose inflation is 11 per cent and nominal interest rates are 15 per cent to keep a pre-tax real interest rate of 4 per cent. Suppose lenders must pay income tax on nominal interest income. The after-tax nominal interest rate is 9 per cent (0.6×15). Subtracting 11 per cent inflation, the after-tax *real* interest rate is -2 per cent. This compares with +2.6 per cent when inflation was zero.

Inflation accounting uses fully inflation-adjusted definitions of costs, income and profit.

When inflation was 11 per cent, nominal interest rates were 15 per cent. Eleven per cent of this was not real income, merely a payment to keep up with inflation. Only 4 per cent was the real interest rate providing real income. But income tax applied to all 15 per cent. Higher inflation reduced the real return on lending because the tax system was not properly inflation-adjusted; that is, it did not use **inflation accounting**. The government gained more real tax revenue. You lost.

Capital gains tax is another example. Suppose people pay tax of 40 per cent on any capital gain made when asset prices rise. When inflation is zero, only real gains are taxed. When inflation is 10 per cent, nominal

asset prices rise merely to preserve their real value. People pay capital gains tax even though they are not making real capital gains.

Institutional imperfections help explain why inflation has real effects even when inflation is fully anticipated. These effects can be large. Usually, the government is the winner.

Unexpected inflation

Previously, we assumed that inflation was fully anticipated. What if inflation is a surprise?

Redistribution

When prices rise unexpectedly, people with nominal assets lose and people with nominal liabilities gain. Nominal contracts to buy and sell, or lend and borrow, can reflect expected inflation, but cannot reflect surprise inflation.

Expecting inflation of 10 per cent, you lend £100 for a year at 12 per cent, expecting a real interest rate of 2 per cent. Unexpectedly, inflation is 20 per cent. The real interest rate on your loan is $[12 - 20] = -8$ per cent.

You lose by lending. Conversely, borrowers gain 8 per cent. Their nominal income rises 20 per cent with inflation but they repay at 12 per cent interest. For every borrower, there is a lender. One person's gain is another person's loss. In the aggregate, they cancel out. But unexpected inflation redistributes real income and wealth; in this case, from lenders to borrowers. This may lead to economic dislocation. Some people may have to declare bankruptcy, which then affects other people. We also have to make a value judgement about whether we like the redistribution that is taking place.

One redistribution is between the government and the private sector. *Unexpected* inflation reduces the real value of all outstanding nominal government debt. It is as if the government had taxed us in order to repay this debt.²

The old and the young

In practice, many savers are the old. Having paid off their mortgages and built up savings during their working life, they put their wealth into nominal bonds to provide income during retirement. These people lose out from surprise inflation.

Nominal debtors are the young and, mainly, those entering middle age with a large mortgage. They gain when surprise inflation raises house prices and nominal incomes without a matching rise in the nominal sum they owe the bank or building society.

Surprise inflation redistributes from the old to the young. We may judge this redistribution undesirable. With technical progress and productivity growth, each generation is richer than the one before. Redistribution from the old to the young raises intergenerational inequality.

Uncertain inflation

Uncertainty about future inflation has two costs. First, it makes planning more complex, raising the real resources society uses to make plans and do business.

Second, people dislike risk. The extra benefits of the champagne years are poor compensation for the years of starvation. People would rather average out these extremes and live comfortably all the time. The psychological costs of worrying about how to cope with the bad years may also be important.

When people make nominal contracts, uncertainty about inflation means uncertainty about the eventual real value of the nominal bargains currently made. This is a true cost of inflation. If a lower average level of inflation also reduces uncertainty about inflation, this may be a reason to aim for low inflation. The institutions that commit the government to low inflation may also reduce the scope for uncertainty about inflation. If so, lower average inflation has a real benefit because it is also more certain.

CASE 22.1

PUBLIC ENEMY NUMBER TWO

Deflation is negative inflation, when the price level is falling.

For several decades, policy makers convinced themselves that inflation was public enemy number one. Inflation is certainly destructive, for the reasons discussed above. However, when financial crisis erupted, concerns about inflation were temporarily but completely set aside. Saving the banking system was more important,

preventing another Great Depression was more important, and getting output on the path to recovery was more important. Several of these judgements reflected a fear that Western economies were about to experience **deflation**.

If inflation is bad, you might be forgiven for thinking that deflation is good. Nothing could be further from the truth. Price stability is good. Low inflation is good. Negative inflation is horrible. Here's why.

Imagine you have borrowed money, and have a nominal debt of £1000. If inflation is foreseen and is 10 per cent, the chances are you had to pay 12 per cent interest in order to provide the lender with a real interest rate of 2 per cent.

Now suppose inflation is -2 per cent. A nominal interest rate of zero will still achieve a real interest rate of 2 per cent. But what happens if inflation is -4 per cent. Nominal interest rates cannot fall below zero, so now the real interest rate is $[(0 - |-4|)] = +4$ per cent. This situation cripples the borrower.

Worse yet, if, as a result of this heavy debt burden, borrowers then spend less, reducing aggregate demand, this puts downward pressure on inflation, taking it to, say, 25 per cent, which causes a bigger debt burden still, a further reduction in aggregate demand, further deflation and yet higher real interest rates. This is an economic black hole, a vicious spiral downwards, from which the economy may not easily escape.

Understanding the dangers of deflation, then, makes sense of two things. First, this is why we normally set inflation targets for monetary policy at 2 per cent not 0 per cent. This provides a margin of safety before any dangers of deflation arise. Aiming on average for zero inflation is a bit too close to the edge of the cliff.

Second, the threat of deflation at the height of the financial crisis was one reason why so many governments loosened fiscal policy at that time. They preferred to cope later with the problem of large government debt rather than cope immediately with debt deflation.

As Yale professor, Irving Fisher, put it during the Great Depression:

In the great booms and depressions [. . .] there were two dominant factors, namely over-indebtedness to start with and deflation following soon after; also that where any of the other factors do

become conspicuous, they are often merely effects or symptoms of these two.³

During 2009, quite a few countries did experience deflation.

Country	Inflation(%)
Canada	-1.9
Denmark	-0.5
Ireland	-3.8
Netherlands	-0.3
Norway	-3.3
Slovakia	-0.6

Source: OECD, Economic Outlook 2012, Statistical Annex.

Without fiscal expansion, the slashing of interest rates and quantitative easing, we might already be stuck in a downward spiral. The countries above avoided protracted deflation partly by their own efforts, and partly because a sufficient number of their trading partners re-established positive growth by stimulatory policy. The more globalized the world, the more a single country's aggregate demand depends on the actions of its partners as well as itself. The prospect of rising export demand provides a backstop in many situations.

Notice that there are now two different reasons why inflation might arise as a postscript to the financial crash. First, governments might be unwilling or unable to raise sufficient tax revenue to service the interest on their huge debt levels. They might create money to finance budget deficits. Second, and nothing to do with the fiscal argument, governments might prefer a period of slightly higher inflation to overcome the fact that, when prices are stable, real interest rates cannot be negative even when nominal interest rates have fallen to zero. This second argument for inflation aims to make monetary policy more powerful by allowing negative real interest rates, which requires positive inflation when nominal interest rates are close to zero.

After two decades of actual or near deflation, following the property crash in the 1990s, Japan finally elected a government in December 2013 committed to *raising* the inflation target in order to break out of the stagnation and deflation.

22.6 Controlling inflation

Policy makers are reluctant to sanction higher inflation because they have spent the last 30 years bringing inflation down. We now discuss how they did it and what they might have to do again. Essentially, inflation is low if people expect it to be low. Credibility is vital.

Incomes policy

Incomes policy is the direct control of wages and other incomes.

A freeze on wage increases certainly gets inflation down quickly. Historically, it has not been able to keep inflation down. Why were past incomes policies unsuccessful?

Once governments intervene in the labour market, they often cannot resist pursuing other aims at the same time. For example, they try to compress relative wages across different skills in the name of fairness. Such policies alter real wages for particular skills, causing excess supply in some skills and excess demand in others. Market forces eventually break the policy.

At best, incomes policy is a temporary adjustment device. In the long run, low nominal money growth is essential if low inflation is to be maintained. Some incomes policies failed because governments introduced a wage freeze but kept printing money – a guarantee that excess demand for workers would eventually break the policy.

Long-term incomes policies are also hard to administer because equilibrium real wages for particular skills change over time. Freezing the existing wage structure gradually sets up powerful market forces of excess supply and excess demand.

Institutional reform

This approach is concerned not with the temporary costs of first getting inflation down, but with how to *keep* inflation down. Central bank independence is a useful pre-commitment to tight monetary policy and low inflation. Here are some examples drawn from the past 20 years.

The Maastricht Treaty

Signed in 1991, the treaty set out conditions both for entering the Eurozone and after admission to it. The first requirement was to avoid loose fiscal policy: a ceiling of 3 per cent on budget deficits relative to GDP.

High-debt countries were also supposed to initiate actions to bring their debt/GDP levels below 60 per cent to minimize the possibility that member states could get into fiscal difficulties and pressurize the European Central Bank to create money to help them out. Moreover, Eurozone entrants first had to succeed in disinflating to low levels, measured both directly by changes in price indexes and indirectly by nominal interest rates.

Not only did EU governments have to sign up for tight policy in the 1990s and beyond, Eurozone hopefuls had to undertake institutional reform, making their national central banks formally independent. The Maastricht Treaty also made the new European Central Bank independent of government, with a mandate to pursue price stability.

It is easy to make commitments, but harder to stick to them, as those in the Eurozone have discovered in the aftermath of the crisis.

Bank of England independence

In May 1997 the new Chancellor, Gordon Brown, gave the Bank of England ‘operational independence’ to set interest rates. The Bank aims to achieve an inflation target set by the Chancellor. In an emergency (a very adverse supply shock), the government can temporarily raise the target rather than force the Bank to initiate a drastic recession merely to hit the inflation target quickly. Nevertheless, any change in the target is politically hard except in truly exceptional circumstances. Operational independence is a commitment to policies favouring low inflation.

ACTIVITY 22.1

IS ASIA’S INFLATION UNDER CONTROL?

In early 2010, Asian inflation rose sharply, after a muted period during the financial crash. Unlike Europe and North America, where fears surrounding renewed recession kept the lid on inflation, Asia was experiencing its highest inflation for over a decade.

India, where prices were flat in mid-2009, had wholesale price inflation of almost 10 per cent; in Vietnam it was 8.5 per cent, in

South Korea 4.2 per cent and in China around 3 per cent. In Australia, with inflation creeping upwards, the Reserve Bank of Australia became the first central bank to increase interest rates several times after the financial crash. Meanwhile, Europe and the US were still pursuing quantitative easing and keeping interest rates as low as they dared.

Indian inflation was largely driven by food and oil prices, while Australian inflation was driven by demand from China for commodities. But why did these supply shocks occur at this particular time? ‘The main driver was the speed and breadth of the recovery, and corresponding aggregate demand generated by most of Asia outside Japan’, concluded the *Financial Times*.⁴ For emerging Asia as a whole, the level of GDP by early 2010 was 4 per cent above its previous peak in the third quarter of 2008, before the output effects of the financial crash then took their toll.

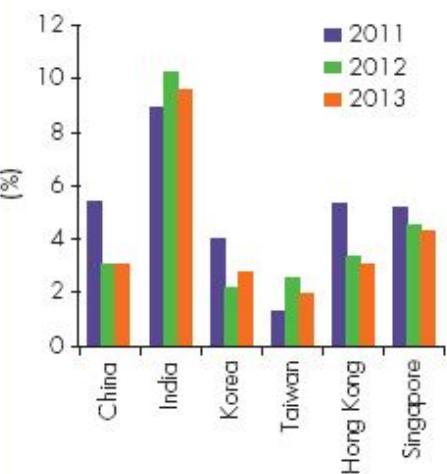
Did this trend continue? The figure shows that inflation remained stubbornly high in India but generally subsided elsewhere. Economic growth in China and India slowed in 2012, taking some of the heat off demand for raw materials and commodity prices.

Even so, by early 2013 new worries about Asian inflation were emerging. Economist Andy Xie told US broadcaster CNBC:

Unprecedented monetary easing by central banks (around the world) will push prices higher. India’s inflation could rise to 10 per cent, and above 5 per cent in Southeast

Asia. Inflation in China could rise to 4 per cent, twice its current level.⁵

Many emerging markets believe that previous bouts of quantitative easing in stagnating developed economies ended up spilling over into inflation in emerging markets, even though lack of demand in the developed world prevented any resurgence of inflation in the countries actually undertaking the quantitative easing. Supporters of QE respond that the principal purpose of QE was to *prevent* a fall in broad money that would otherwise have occurred when bank deposit multipliers collapsed, rather than to *increase* the supply of broad money.



Inflation in emerging Asia, 2011–13

Source: IMF, *World Economic Outlook*, November 2012.

Questions

- If Europe, North America and Japan were still stagnating, what must have been the source of aggregate demand for Asian goods?
- How much spare capacity – for output and for labour – is likely now to be left in emerging Asian markets?
- What does this imply about the likely course of future monetary policy in these countries?
- After two decades of stagnation, Japan remains concerned about deflation. Would you expect Japanese monetary policy now to follow policies in Europe and the US, or those in China and India?
- If energy, food and commodities prices rise sharply, what dilemma will then face policy makers in Europe, the US and Japan?

To check your answers to these questions, go to page 681.

22.7

The Monetary Policy Committee

Underlying inflation is the growth of the retail price index (RPI), after omitting the effect of mortgage interest rates on the cost of living (hence the abbreviation RPIX).

Headline inflation is actual inflation, the growth in the RPI.

Since 1997 UK interest rates have been set by the Bank of England's Monetary Policy Committee (MPC), which meets monthly to set interest rates to try to hit the inflation target laid down by the Chancellor. Initially, the target was 2.5 per cent annual inflation, plus or minus 1 per cent. The target applied to **underlying inflation**(which ignores mortgage interest rates) not **headline inflation**.

Why omit mortgage interest from the price level on which monetary policy should focus? Suppose inflation is too high. To reduce aggregate demand, interest rates are raised. But higher interest rates *raise* the RPI by raising the cost of living for homeowners. Moreover, when temporary changes in interest rates are required to get the economy back on track, it may also be more sensible to target the underlying rate of inflation.

Different countries construct price indexes in slightly different ways. EU countries CPI inflationmeasures the rate of growth of an index of consumer prices.have each adopted a common procedure for calculating their consumer price index (CPI), making cross-country comparisons of inflation more meaningful. In December 2003, the UK Chancellor, Gordon Brown, instructed the Bank of England to switch from using the RPIX to using the CPI as the basis for inflation targeting. For statistical reasons, **CPI inflation**tends to grow less rapidly than RPIX inflation. At the time of the crossover, UK inflation was 2.9 per cent measured by the growth rate of the RPIX but only 1.3 per cent measured by the growth rate of the CPI. Hence, Gordon Brown also changed the target inflation rate from 2.5 per cent growth in the RPI to 2.0 per cent growth in the CPI.

CPI inflation measures the rate of growth of an index of consumer prices.

The quarterly *Inflation Report* includes the famous **fan chart**for CPI inflation. Figure 22.9 shows the fan chart for November 2012. A darker projected line implies a higher probability of that inflation outcome. Figure 22.9 shows that, in November 2012, the Bank was expecting UK inflation to average around 2 per cent in 2013, and then perhaps increase a little thereafter as global growth resumed more strongly.

A **fan chart** indicates the probability of different outcomes.

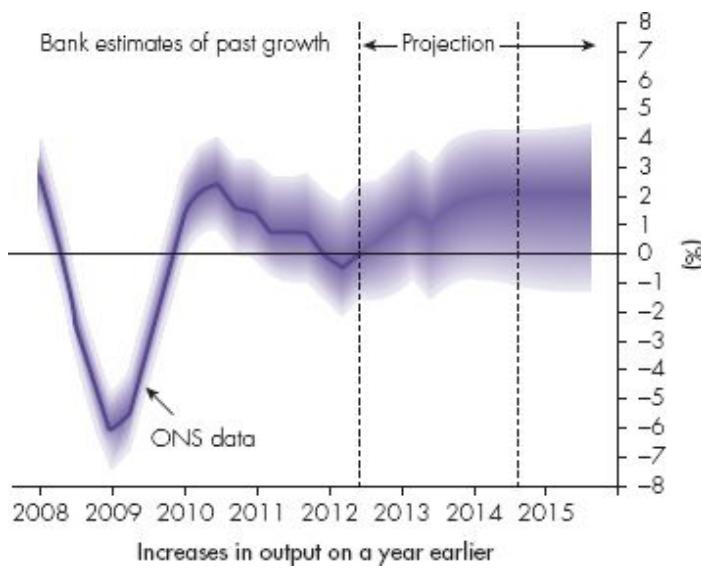


Figure 22.9 UK inflation projection made in November 2012

Source: © Bank of England, taken from Inflation Report November 2012,
www.bankofengland.co.uk/publications/pages/inflationreport

The Bank has been criticized for its track record of under-predicting UK inflation, and it is not impossible that this will happen again. If there is a boom in commodity prices (more demand, driven by the Asian boom) and food prices (lower supply, caused by floods, droughts and other implications of climate change), then UK monetary policy will face a very tough choice. Does it raise interest rates to choke off imported inflation, despite a feeble economic recovery, or does it keep interest rates low and allow inflation to exceed the target yet again?

In this section, we discuss three questions. Why was the MPC given a target for inflation? How does it work? How easy was it for the MPC to decide where to set interest rates?

Inflation targets

Without a nominal anchor, nothing ties down the price level or any other nominal variable. Market forces only determine real variables.

Nominal money is a possible nominal anchor and is attractive as an intermediate target because new data on money come out faster than data on prices or output. Monetary targets fell out of favour because large and unpredictable changes in real money demand made it hard to know where

to set the nominal money target; more recently, broad money supply has also been unpredictable because of changes in the bank deposit multiplier. When it is hard to predict M/P , it is hard to know where to set M in order to get the desired path of P .

As explained in previous chapters, most modern central banks implicitly follow a Taylor rule but their policy is often portrayed and communicated as a flexible inflation target. This is easily understood by the public, and more easily monitored than a Taylor rule, which could lead to disputes about what the (unobservable) level of potential output really is.⁶

Back to the future

Delays in data availability mean that the MPC has to forecast where the economy is today. Moreover, the interest rate medicine takes up to two years to have its full effect on private behaviour. Hence the MPC has to *forecast* the path of prices at least two years into the future merely to know where to set interest rates *today*!

On occasion, the MPC may raise interest rates even though current inflation is under control. This means that, in the absence of any change in interest rates, the MPC is forecasting that inflation will be too high. It then has to act quickly to keep inflation on track.

Reasonable on inflation, shame about the crash

Despite overshooting its inflation target by small amounts, the MPC successfully maintained UK inflation within a much narrower range than previously accomplished. The Bank was prepared to change interest rates even when this was unpopular. But low levels of inflation led to low nominal interest rates, which encouraged more reckless private sector behaviour. Determining sensible monetary policy was never the entire remit of the Bank of England – it has a *financial stability* wing as well as a *monetary stability* wing – and the Financial Services Authority was explicitly charged with financial regulation. Good monetary policy cannot be held responsible for inadequate financial regulation.

Figure 22.10 shows the history of UK interest rates since 1980. Although the Bank's operational independence to set interest rates was formally granted in 1997, Figure 22.10 shows that the decisive break was in 1992 when sterling left the Exchange Rate Mechanism and changed nominal anchors from a pegged exchange rate to an inflation target. Reinforced by

formal independence, since 1997 the MPC has built on its earlier success during 1992–97. The low interest rates since 2008 are clearly an abnormal response, caused by the unique period since the financial crash.

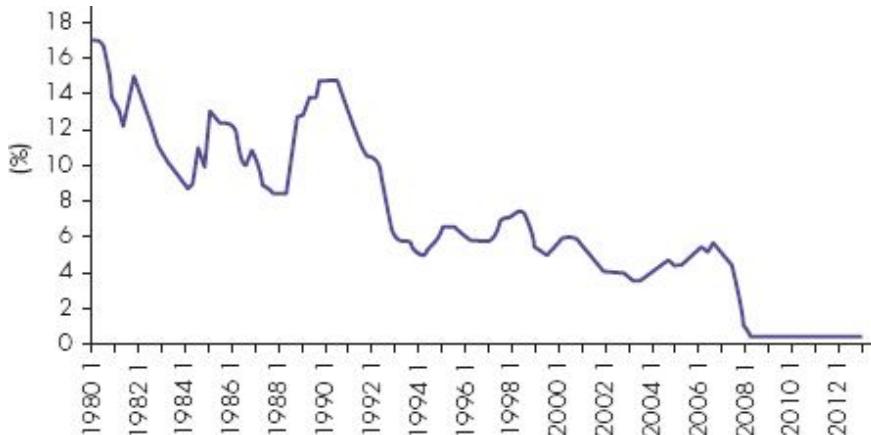


Figure 22.10 UK interest rates, 1980 Q1–2013 Q1 (%)

Source: www.bankofengland.co.uk/monetarypolicy.

Summary

- The **quantity theory of money** says changes in prices are caused by equivalent changes in the nominal money supply. In practice, prices cannot adjust at once to changes in nominal money, so interest rates or income alter, changing real money demand. Nevertheless, in the long run, changes in prices are usually associated with changes in nominal money.
- The **Fisher hypothesis** is that a 1 per cent rise in inflation leads to a similar rise in nominal interest rates so real interest rates change little. Since the nominal interest rate is the cost of holding money, higher inflation reduces real money demand. The *flight from cash* during hyperinflation is a vivid example.
- For a solvent government, there need be no close relationship between the budget deficit and nominal money growth. In the long run, persistent borrowing to finance large deficits may leave the government so

indebted that further borrowing is impossible. It must resort to printing money or take fiscal action to cut the deficit.

- The **long-run Phillips curve** is vertical at equilibrium unemployment. If people foresee inflation and can completely adjust to it, inflation has no real effects.
- The **short-run Phillips curve** is a temporary trade-off between unemployment and inflation in response to demand shocks. Supply shocks shift the Phillips curve. The height of the short-run Phillips curve also depends on underlying money growth and expected inflation. The Phillips curve shifts down if people believe inflation will be lower in the future.
- Temporary supply shocks also shift the short-run Phillips curve.
Stagflation is high inflation plus high unemployment.
- Some so-called **costs of inflation** reflect inflation illusion or a failure to see inflation as the consequence of a shock that would have reduced real incomes in any case. The true costs of inflation depend on whether it was anticipated and on the extent to which the economy's institutions allow complete inflation-adjustment.
- **Shoe-leather costs** and **menu costs** are unavoidable costs of inflation and are larger the larger the inflation rate. Failure fully to inflation-adjust the tax system may also impose costs, even if inflation is anticipated.
- **Unexpected inflation** redistributes income and wealth from those who have contracted to receive nominal payments (lenders and workers) to those who have contracted to pay them (firms and borrowers).
- Uncertainty about future inflation rates imposes costs on people who dislike risk. Uncertainty may be greater when inflation is already high.
- **Incomes policy** may accelerate a fall in inflation expectations, allowing disinflation without a large recession. But it is unlikely to succeed in the long run. Only low money growth can deliver low inflation in the long run.

- **Operational independence of central banks** is designed to remove the temptation faced by politicians to print too much money.

Review questions



EASY

- 1 Equal annual payments in nominal terms become declining annual payments in real terms. Does this explain why voters mind high inflation even when nominal interest rates rise in line with inflation?
- 2 Looking at data on inflation and unemployment over ten years, could you tell the difference between supply shocks and demand shocks?
- 3 **Common fallacies** Why are these statements wrong? (a) Getting inflation down is the only way to cure high unemployment. (b) Inflation stops people from saving. (c) Inflation stops people from investing.
- 4 Which of the following statements is correct? (a) The long-run Phillips curve should really have a positive slope because higher inflation makes firms substitute away from workers who are causing the underlying problem. (b) If inflation leads people to economize on some forms of money, this must make the economy less productive and probably raises long-run unemployment. (c) When other things are assumed to be equal, it is a tolerable approximation to view the long-run Phillips curve as vertical.

MEDIUM

- 5 Name three groups which lose out during inflation. Does it matter whether this inflation was anticipated?
- 6 (a) Explain the following data taken from *The Economist* a few years ago (when some countries still had proper inflation!). (b) Is inflation always a monetary phenomenon?

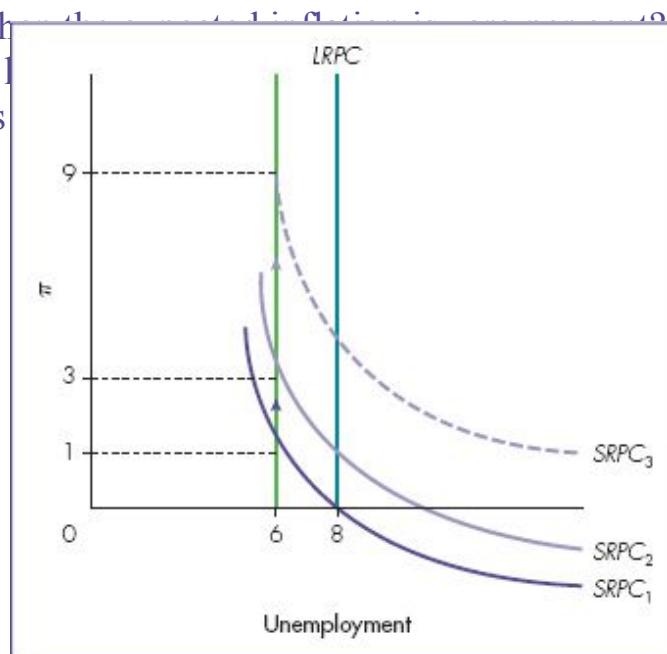
	Money growth (%)	Inflation (%)
Eurozone	3	2
Japan	12	-3
UK	6	2
Australia	15	3
US	8	2

- 7 Professor Milton Friedman argued that money was socially useful but essentially free to create. Society should therefore reduce the opportunity cost of holding money to zero, so that people would demand it up to the point at which its marginal benefit was zero. (a) Suppose the real interest rate on other assets is around 3 per cent. Is there any way society could arrange for cash to earn a similar real return? (b) Why don't governments do this?
- 8 Your real annual income is constant, and initially is £10 000. You borrow £200 000 for 10 years to buy a house, paying interest annually and repaying the £200 000 in a final payment at the end. (a) List your annual incomings and outgoings in the first and ninth year if inflation is 0 and the nominal interest rate is 2 per cent a year. (b) Repeat the exercise if annual inflation is 100 per cent and the nominal interest rate is 102 per cent. Is the real interest rate the same in both situations?
- 9 Inflation in Zimbabwe, high for many years, reached hyperinflation levels in the recent past. (a) President Mugabe blamed Western governments for restricting trade and driving up prices. Could a fall in supply have generated sustained high inflation? (b) Why do you think Zimbabwe has such high inflation? (c) Is inflation high enough to raise the maximum possible revenue for the government?
- 10 Suppose D is real government debt, s the primary budget surplus $T - G$ (that is, excluding interest payments on debt), i the real interest rate, Y real output and g the rate of output growth. The debt burden D/Y rises with debt but falls with output and the ability to repay debt. Let Δ denote the increase in a variable. (a) If $\Delta(D/Y) = (\Delta D/D) - (\Delta Y/Y)$, show that the debt/GDP ratio shrinks only if $s/D > i - g$. (b) Suppose all debt is cash, paying no interest. Show that the above relationship becomes $s/D > (g + \pi)$.
- 11 Suppose Asia emerges from the financial crisis much more quickly than Europe and North America. As China and India bid up world commodity prices, use a figure resembling Figure 22.8 to illustrate the effect on European unemployment.

HARD

- 12 Draw a curve to illustrate how the real revenue raised by the government through foreseen inflation varies with the inflation rate. If an economy moves from using a lot of cash to using a lot of electronic money on which market interest rates are paid, illustrate how the curve changes.
- 13 The diagram below illustrates the short-run and long-run Phillips curves. (a) Why is the long-run Phillips curve vertical in the diagram? (b) What

is the rate of unemployment which
(c) Use the diagram to explain how the
unemployment rate as people's



[4 Essay question] Does the huge success of central bank independence in so many countries suggest that other decisions should be removed from government? Your answer should include assessments of the case for (a) an independent health services board, (b) an independent budget deficit commission, and (c) a redistribution commission.

- 1 An exception is the hyperinflation example of the next section.
- 2 Why stress unexpected inflation? Because expected inflation is already built into the terms on which bonds were originally issued. Expected inflation affects nominal interest rates.
- 3 Irving Fisher, 'Debt-deflation theories of great depressions', *Econometrica* 1, no. 4 (1933): 337–357.
- 4 'Asia's inflation genie leaps out of the bottle', *Financial Times*, 18 March 2010.
- 5 *Watch out, Asia: Inflation is coming*, 2 January 2013 (www.cnbc.com).
- 6 Like central banks deciding where to set interest rates, academic researchers engaged in empirical evaluation of monetary policy have to make estimates of how potential output is evolving. The OECD regularly publishes estimates of output gaps ($Y - Y^*$) for the major countries.

CHAPTER 23

Unemployment

Learning outcomes

By the end of this chapter, you should be able to:

- 1 define classical, frictional and structural unemployment
- 2 distinguish between voluntary and involuntary unemployment
- 3 discuss measured unemployment, both claimant count and standardized rate
- 4 analyse determinants of unemployment
- 5 explain how supply-side policies reduce equilibrium unemployment
- 6 evaluate private and social costs of unemployment
- 7 explain hysteresis

In the early 1930s, nearly a quarter of the UK labour force was unemployed, and other countries suffered similarly. Society threw away output by failing to put people to work. For the next 40 years macroeconomic policy tried to manage aggregate demand to avoid a rerun of the 1930s. Figure 23.1 shows that until the 1970s the policy succeeded.

In the 1970s high inflation emerged for reasons discussed in the previous chapter. Governments eventually tightened monetary and fiscal policy to get inflation under control. The mix of tighter demand policies and adverse supply shocks led to a big rise in unemployment in the 1980s.

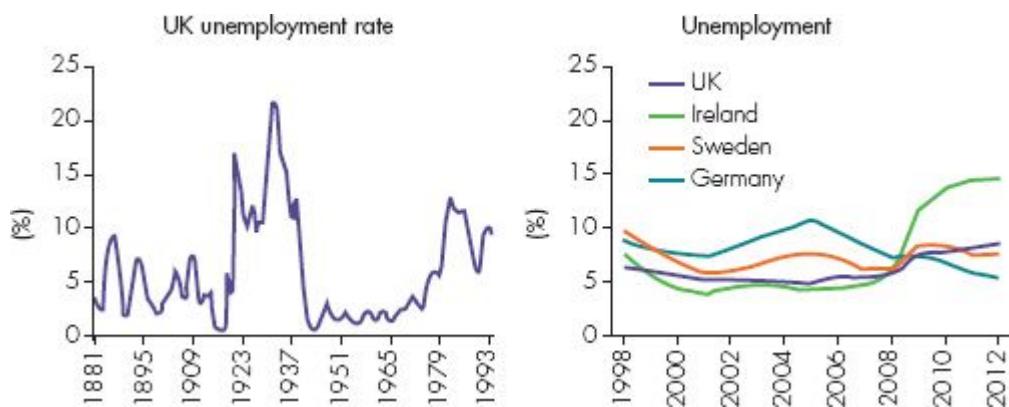


Figure 23.1 Unemployment since 1881

Source: 'Unemployment statistics from 1881 to the present day,' *Labour Market Trends*, January 1996; OECD, *Economic Outlook, Statistical Annex 2012*.

After the economy adjusted, deficient demand was no longer the cause of high unemployment. Equilibrium unemployment remained high because of adverse changes in supply. Better supply-side policies since the mid-1990s reduced UK unemployment to levels not seen since before the 1970s.

The right-hand panel also shows recent data for Germany, Ireland and Sweden. Although, like the UK, these countries experienced unemployment in the 5–10 per cent range in the years before the financial crash, their subsequent experiences differed. In order to secure a bailout, the Irish government had to adopt drastic austerity which drove Irish unemployment back up to almost 15 per cent, much higher than the level in the UK, Sweden or Germany. By 2013, Spanish unemployment (not shown above) had reached 27 per cent.

23.1 The labour market

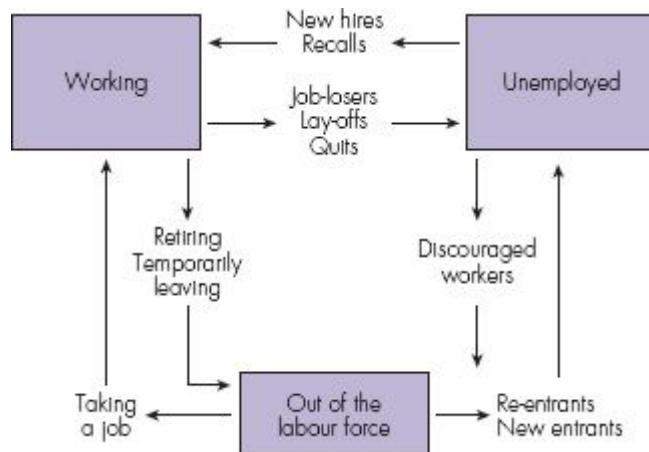
Not everyone wants a paid job. Full-time homemakers, voluntary workers, the old and the young, and those with long-term sickness are all excluded from the category of people seeking paid work. The people who want a job are called the **labour force**. The **participation rate** measures how many people of working age want to work.

The **labour force** comprises people with a job or registered as looking for work at the current wage rate.

The **participation rate** is the fraction of the population of working age in the labour force.

The **unemployment rate** is the fraction of the labour force without a job but registered as looking for work.

Some people looking for work do not register as unemployed. They do not appear in official statistics for the registered labour force or the registered unemployed. Yet from an economic viewpoint, such people *are* in the labour force and *are* unemployed. For the moment, our data on the labour force or the unemployed refer only to those registered. The **unemployment rate** measures unemployment relative to the size of the labour force.



A person may be working, unemployed or out of the labour force. The arrows show the routes along which people move. Each route carries a surprisingly heavy flow.

Figure 23.2 Labour market flows

Figure 23.1 showed that UK unemployment was high in the inter-war years, especially in the 1930s. By comparison, the post-war unemployment rate was tiny until the late 1970s. In the 1980s it started to get back to pre-war levels, but then fell thereafter. Was this because policy makers learned the secret of demand management or because structural reforms in the labour market reduced equilibrium levels of

unemployment? By the end of this chapter, you should have a clearer idea about the answers.

Stocks and flows

Unemployment is a stock concept measured at a point in time. Like a pool of water, its level rises when inflows (the newly unemployed) exceed outflows (people getting new jobs or quitting the labour force altogether). Figure 23.2 illustrates this important idea.

There are three ways for workers to become unemployed. Some people are sacked or made redundant (job-losers); some are temporarily laid off but expect eventually to be rehired by the same company; and some voluntarily quit their existing jobs. But the inflow to unemployment also comes from people not previously in the labour force: school-leavers (new entrants) and people who, having left the labour force, are now returning to look for a job (re-entrants).

People leave the unemployment pool in the opposite direction. Some get jobs. Others give up looking for jobs and leave the labour force completely. Some of this latter group may simply have reached the retirement age at which they get a pension, but many are **discouraged workers**.

Table 23.1 shows that the pool of unemployment is not stagnant. In 2011 Q3, of UK adults between the ages of 16 and 64, there were 2.7 million unemployed, 9.2 million who were inactive (outside the labour market) and 28.3 million in employment. Figure 23.3 shows that there had been considerable two-way flows between each of these categories since the previous quarter 2011 Q2. In fact, in a single quarter, 2.9 million people were moving categories within the labour market, a larger number than the entire stock of unemployment. The labour market is not stagnant.

Table 23.1 Unemployment by duration, October 2012 (million people)

< 6 months	6–12 months	12–24 months	24 + months
1.2	0.4	0.4	0.5

Source: ONS, *Labour market statistics*, December 2012.

When unemployment is high, people often have to spend longer in the pool before they find a way out. Table 23.1 gives data on the duration of unemployment. Unemployment is not always a temporary stopover on the way to better things. A higher unemployment rate usually also means

that people are spending longer in the pool of unemployment before escaping. Table 23.1 shows the 2.5 million UK unemployed in October 2012, according to the period for which they have already been unemployed.

Even by late 2012, the labour market had not experienced the massive rise in unemployment that the sharp 2009 output contraction would usually have implied. Why did unemployment not increase by even more than it did? Since labour is expensive both to recruit and to dismiss, firms try to ride out a temporary storm by hoarding the workers that they have. Once the recession is evidently going to persist for some time, firms then have little choice but to adjust the number of workers in their employment. In Table 23.1 there are relatively few ‘long-term’ unemployed because the previous decade had been a period of economic growth with strong employment opportunities. The longer the recession persists, the more we expect to see a ‘drift to the right’ in Table 23.1, with larger numbers out of work for longer periods at a stretch.

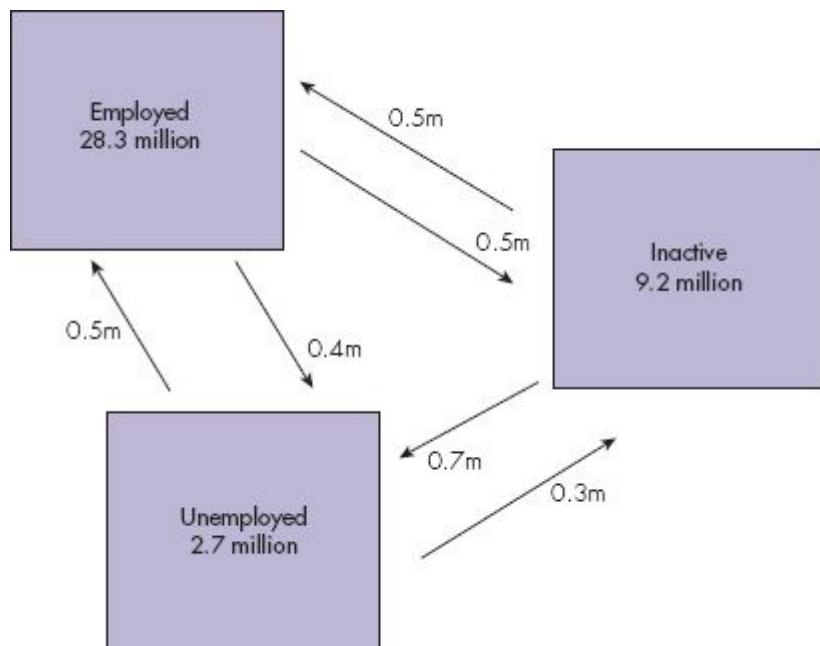


Figure 23.3 Labour market flows, 2011 Q3 (million people per quarter)

Source: www.ons.gov.uk/ons/.

It may also be the case that changes in the structure of the labour market, and in technology, have enabled much greater part-time working than would have been the case in previous recessions. Although the demand for labour contracted, some of this was met by working shorter hours rather than by throwing people entirely out of work. The Internet has made self-employment and flexible working from home much easier than before.

CASE 23.1

MEASURING UNEMPLOYMENT

The unemployed are those without a job but willing to work at the prevailing wage rate – the difference between the labour force and those with jobs. Measuring those with jobs is not so controversial, but how do we measure the labour force? How do we know how many people would like to work at the current wage rate?

Claimant unemployment

One convenient short cut is simply to measure the number of people in receipt of unemployment-related benefit. In the UK, this transfer payment from government used to be called unemployment benefit. To diminish any stigma attached, it was renamed jobseeker's allowance.

Over time and across countries, governments differ in the generosity of the eligibility criteria for claiming this benefit. If we use the claimant count as our measure of unemployment, any government attempt to toughen the criteria for eligibility for benefit will appear to have reduced measured unemployment, and international comparisons are hard to interpret because of national differences in eligibility for benefit.

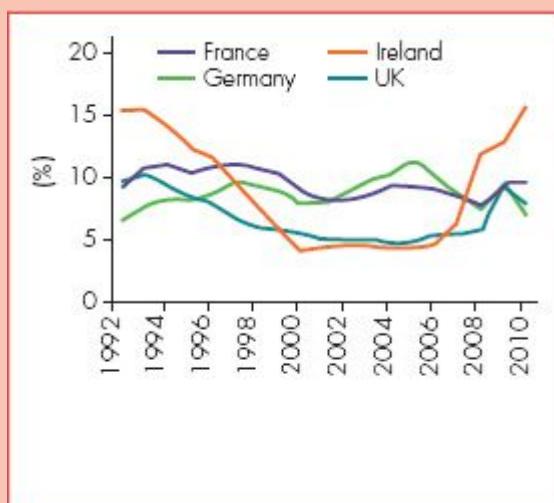
Those ineligible for UK jobseeker's allowance included: (a) those with savings above a certain amount, (b) those unable to work at least 40 hours a week, (c) those aged 16–17, and (d) those unwilling to sign a 'jobseeker's agreement' systematically to seek work. Those whose previous job terminated because of misconduct did not have an automatic right to jobseeker's allowance. Other countries make other stipulations. Those ineligible for unemployment-related

benefits are usually eligible for a residual welfare safety net at a less generous level.

Thus, in every country, measured unemployment based on counting claimants for unemployment-related benefit understates 'true unemployment' because some of those declared ineligible are in fact in the labour force, seeking a job, and unemployed.

Standardized unemployment

Fully aware of this issue, international agencies have endeavoured to produce a more comprehensive definition capable of being compared across countries. Accurate measures of unemployment depend on knowing people's intentions. It would be prohibitively expensive to ask every individual and assess whether their answers were truthful.



Standardized unemployment 1992–2010 (% of labour force)

Source: OECD, *Economic Outlook*, 2012.

Standardized measures use surveys of smaller numbers of people and then extrapolate the answers to an estimate for the entire economy, just as opinion poll surveys try to predict election results and crime surveys try to estimate national crime statistics. Surveys are always subject to a margin of error, but the widespread use of standardized unemployment measures suggests that there is a degree of confidence in this approach.

The surveys ask people of working age, but without work, whether they: (a) are available to work within two weeks and actively job-hunting, (b) are waiting to take up a job already offered or (c) have no wish to have a job at current wage rates. Those replying yes to

(a) and (b) are included in the standardized measure of unemployment.

The figure shows the evolution of standardized unemployment rates during 1992–2010 reported by the OECD. The year 2008 may yet prove to have been a point of low unemployment – as European economies are driven to fiscal austerity to combat their high levels of government debt, they will be delighted if other sources of aggregate demand increase sufficiently quickly to keep pace with the underlying growth of aggregate supply. By 2012, the UK had already experienced a 'double dip recession' and commentators were starting to discuss the prospect of a 'triple dip'.

The composition of unemployment

Table 23.2 gives a recent breakdown of unemployment by gender and age. Young workers find it much harder to get a job. Unlike established workers with accumulated skills and job experience, young workers have to be trained from scratch. Youth unemployment was over 30 per cent, which considerably exceeds the national average of around 8 per cent.

The unemployment rate is lower for women than for men, perhaps because more women leave the labour force if they do not succeed in getting jobs.

Table 23.2 UK unemployment rates, October 2012 (% of relevant group)

Age	Men	Women
16–17	39	34
18–24	21	15
16–64	8	7

Source: ONS, *Labour market statistics*

23.2 Analysing unemployment

We now develop a theoretical framework in which to analyse unemployment. We can classify unemployment by the source of the problem or by the nature of behaviour in the labour market.

Types of unemployment

Frictional unemployment includes people whose handicaps make them hard to employ. More importantly, it includes people spending short spells in unemployment as they hop between jobs in a dynamic economy.

Frictional unemployment is the irreducible minimum unemployment in a dynamic society.

As the labour force evolves, and the skill needs of employers change with new products, services and technologies, finding more effective matches between what workers can offer and what firms require is a significant part of remaining efficient and sustaining productivity growth. Since perfect matching cannot be achieved instantaneously, it is important to recognize that some unemployment is actually good for the economy if it corresponds to a vigorous search for better job matching. Frictional unemployment is sometimes called search unemployment.

Structural unemployment reflects the time taken to acquire human capital. A skilled steelworker may have worked for 25 years but is made redundant at age 50 when the industry contracts in the face of foreign competition. That worker may have to retrain in a new skill which is more in demand in today's economy. Firms may be reluctant to take on and train older workers who have only a short remaining working life in which to repay the expensive investment. Such workers are victims of structural unemployment.

Structural unemployment arises from the mismatch of skills and job opportunities as the pattern of demand and supply changes.

Structural unemployment implies that the market price for a worker's skill has fallen because of a shift in supply or demand, often caused by new technology or changes in international competition. Faced with a job offer at a low wage, the worker may prefer to retrain in order to earn something closer to the previous wage; if there are few such opportunities, the worker may become unemployed. Even low-skilled, low-paid alternatives may be difficult to find. Many such jobs are now outsourced abroad, and domestic service employers may prefer younger workers with fewer demands.

A third type of unemployment is purely the consequence of a fall in aggregate demand. Until wages and prices have adjusted to their new long-run equilibrium level, a fall in aggregate demand reduces output and employment. Some workers want to work at the going real wage rate but cannot find jobs. Only when demand has returned to its long-run level is **demand-deficient unemployment** eliminated.

Demand-deficient unemployment occurs when output is below full capacity.

Classical unemployment describes the unemployment created when the wage is deliberately maintained above the level at which the labour supply and labour demand schedules intersect.

If wages and prices were perfectly flexible, demand-deficient unemployment would not occur. Since the **classical** model assumes that flexible wages and prices maintain the economy at full employment, classical economists had difficulty explaining high unemployment in the 1930s. Within their paradigm, it had to be attributed either to frictional or structural unemployment. They concluded that the wage was prevented from adjusting to its equilibrium level, caused either by the exercise of trade union power or by minimum wage legislation which enforces a wage in excess of the equilibrium wage rate. Effectively, these had created unnecessary structural unemployment.

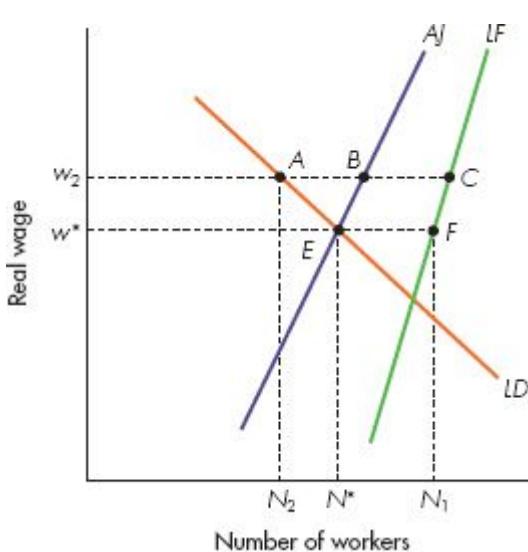
The modern analysis of unemployment takes the same types of unemployment but classifies them differently to highlight the behavioural implications and consequences for government policy. Modern analysis stresses the difference between *voluntary* and *involuntary* unemployment.

Equilibrium unemployment

Figure 23.4 shows the labour market. The labour demand schedule LD slopes down. Firms demand more workers at a lower real wage because the cost of labour is lower. As in most economic applications, demand curves slope downwards.

The LF schedule shows the number of people in the labour force. A higher real wage increases the number of people wishing to work. This is not as obvious as it may at first appear. A higher wage increases the

benefit of an hour of work relative to an hour of leisure. The substitution effect leads to a desire to work more and consume less leisure. But there is also an income effect. A higher wage makes people richer, raising the quantity of goods and leisure demanded. This income effect makes people less interested in working when wages are higher. Figure 23.4 shows what we have learned from considerable empirical research – higher real wages do increase the size of the labour force, but by only a little. The *LF* schedule is pretty steep.



The schedules *LD*, *LF* and *AJ* show, respectively, labour demand, the size of the labour force and the number of workers willing to accept job offers at any real wage. *AJ* lies to the left of *LF* both because some labour force members are between jobs and because optimists are hanging on for an even better job offer. When the labour market clears at *E*, *EF* is the natural rate of unemployment, the people in the labour force not prepared to take job offers at the equilibrium wage w^* . If union power succeeds in maintaining the wage w_2 in the long run, the labour market will be at *A*, and the natural rate of unemployment *AC* now shows the amount of unemployment chosen by the labour force collectively by enforcing the wage w_2 .

Figure 23.4 Equilibrium unemployment

The schedule *AJ* shows how many people accept job offers at each real wage. The schedule is to the left of the *LF* schedule: only people in the labour force can accept a job. Hence, at any horizontal level corresponding to a particular market wage, the *AJ* schedule must lie to the left of the *LF* schedule. How far *AJ* lies to the left of *LF* depends on several things. Some people are inevitably between jobs at any point in time. Also, a particular real wage may tempt some people into the labour

force even though they will accept a job offer only if it provides a higher real wage than average.

We draw these schedules for a given level of jobseeker's allowance. When wages are high, jobseekers grab available jobs. The two upward-sloping schedules are close together. When wages are low (relative to unemployment benefit), potential workers are more selective in accepting job offers. People invest in searching longer for an even better job. The two schedules are further apart.

Labour market equilibrium is at E in Figure 23.4. *Equilibrium employment* is N^* . The distance EF is **equilibrium unemployment**. This unemployment is entirely *voluntary*.

Equilibrium unemployment (also called the natural rate of unemployment) is the unemployment rate when the labour market is in equilibrium.

A worker is **voluntarily unemployed** if, at the given level of wages, she wishes to be in the labour force but does not yet wish to accept a job.

At the equilibrium real wage w^* , N_1 people want to be in the labour force but only N^* accept job offers; the remainder do not yet want to work at the equilibrium real wage.

Equilibrium unemployment includes frictional and structural unemployment. Suppose a skilled welder earned £1000 a week before being made redundant. The issue is not why workers became redundant (the decline of the steel industry), but why these workers will not take a lower wage as a dishwasher to get a job. Their old skills are obsolete. Until new skills are learned, dishwashing may be their only skill valued by the labour market. People not prepared to work at the going wage rate for their skills, but wanting to be in the labour force, are **voluntarily unemployed**.

A worker **involuntarily unemployed** would accept a job offer at the going wage rate.

First time round, it can be difficult to grasp the concept of voluntary unemployment. The key is to remember that, at any instant, we are taking a snapshot of a highly dynamic labour market. Take another look at Figure 23.4. The gross number of people flowing between employment, unemployment and inactivity (being out of the labour force) is larger every quarter than the total stock of unemployment. With so many people in motion, at any particular measurement point they cannot all be recorded as employed or inactive. A voluntarily unemployed worker is interested enough in finding a job to be in the labour force, but has not yet found a job offer that she is prepared to accept. Such workers are still looking and still hoping.

What about classical unemployment, for example if unions keep wages above their equilibrium level? This is shown in Figure 23.4 as a wage w_2 above w^* . Total unemployment is AC . As individuals, AB workers want jobs at the wage w_2 but cannot find them. Firms wish to be at point A . As individuals, the workers AB are **involuntarily unemployed**.

However, through their unions, workers collectively opt for the wage w_2 above the equilibrium wage, thus reducing employment. For workers as a whole, the extra unemployment is voluntary – they could have chosen a different wage and a different employment level, but preferred not to. We include classical unemployment in equilibrium unemployment. If unions maintain the wage w_2 , the economy stays at A and AC is equilibrium unemployment.

CONCEPT 23.1

THE LUMP-OF-LABOUR FALLACY

Those without economics training often think there is a simple solution for reducing unemployment: shorten the working week, so that the same amount of total work is shared between more workers, leaving fewer people unemployed. What's wrong with this argument?

It presumes the demand for labour (hours \times people) is fixed, whatever the cost of hiring workers or their benefit in goods produced and revenue earned. In practice, both would be affected by the proposal.

You go to work for seven hours a day, but probably have an hour of dead time (visiting a coffee shop, tidying your desk, being nice to colleagues, talking about sport, texting friends). This is a fixed cost, say an hour of time. There are probably economies of scale to shift length. Shortening the shift length adds to the cost of labour, making firms less competitive. For any given output demand for their product, from which we can derive the labour demand curve, a higher cost of labour makes firms choose to demand fewer workers. Firms move up their downward-sloping demand curve for labour and offer fewer jobs.

Few economists think compulsory reductions in the length of the working week are a promising solution to the problem of high unemployment.

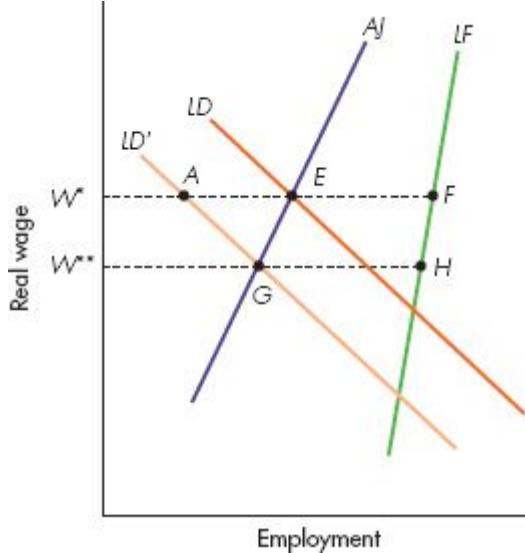
Figure 23.5 illustrates how Keynesian or demand-deficient unemployment may arise. Initially, labour demand is LD and the labour market is in equilibrium at E , with equilibrium unemployment EF . Then labour demand shifts down to LD' . Before wages or prices adjust, the real wage is still w^* . At this wage, workers want to be at E but firms want to be at A . The distance AE is demand-deficient unemployment; that is, involuntary unemployment caused by sluggish adjustment of wages and prices. EF remains voluntary unemployment.

If labour demand remains LD' , eventually real wages fall to w^{**} to restore equilibrium at G . However, by reducing interest rates, monetary policy can shift labour demand up to LD again and restore equilibrium at E . At A , output and employment are low. Involuntary unemployment also reduces wage growth and inflation.

Thus, we can divide total unemployment into two parts. The equilibrium or natural rate is the equilibrium unemployment determined by normal labour market turnover, structural mismatch, union power and incentives in the labour market. Keynesian unemployment, also called demand-deficient or cyclical unemployment, is involuntary unemployment in disequilibrium, caused by low aggregate demand and sluggish wage adjustment.

This division helps us think clearly about the policies needed to tackle unemployment. Keynesian unemployment reflects spare capacity and wasted output. By boosting labour demand, policy can mop up this spare

capacity and increase output and employment. Wage adjustment could logically accomplish the same outcome, but may take several years to do so. The more sluggish are market forces, the more it makes sense for policy to intervene. Most forms of monetary policy have the consequence that interest rates will adjust to such a situation and help offset the original demand shock. The automatic fiscal stabilizers also act in this direction.



Beginning from equilibrium at E labour demand falls from LD to LD' . Before price and wage adjustment occurs, the economy moves to A . EF is still voluntary unemployment, but now AE is involuntary unemployment, since workers want to be at E at a real wage W^* . If labour demand remains LD' , eventually real wages fall to W^* to restore equilibrium at G . By reducing interest rates, monetary policy can shift labour demand up to LD and restore equilibrium at E . Eliminating spare capacity AE allows higher output and employment.

Figure 23.5 Keynesian unemployment

In marked contrast, when the economy is already in long-run equilibrium, further demand expansion is pointless. Even though unemployment is not zero, there is no spare capacity. At points E or G in Figure 23.5, all remaining unemployment is voluntary.

It is true that, beginning from G , shifting labour demand up from LD' to LD achieves a small reduction in equilibrium unemployment. The distance EF is smaller than GH because the AJ and LF schedules are not parallel to one another. However the main effect of raising demand is to bid up wages, not to increase output or employment.

Hence, when the economy begins with only voluntary unemployment, reductions in unemployment and increases in output are mainly accomplished not by demand policies but by supply-side policies. These policies either *shift* the supply schedules *AJ* and *LF* or they reduce distortions that prevented the economy getting to points like *E* or *G*.

The next section presents some evidence on the relative magnitude of unemployment responses to demand and supply, and then analyses these supply-side policies in more detail.

23.3 Explaining changes in unemployment

Empirical research aims to decompose causes of unemployment into those that changed equilibrium and those that caused demand-deficient unemployment. Figure 23.6 compares the actual unemployment rate with estimates of the equilibrium unemployment rate from 1956 to 2009. Averaging data within sub-periods reduces the influence of short-term fluctuations.

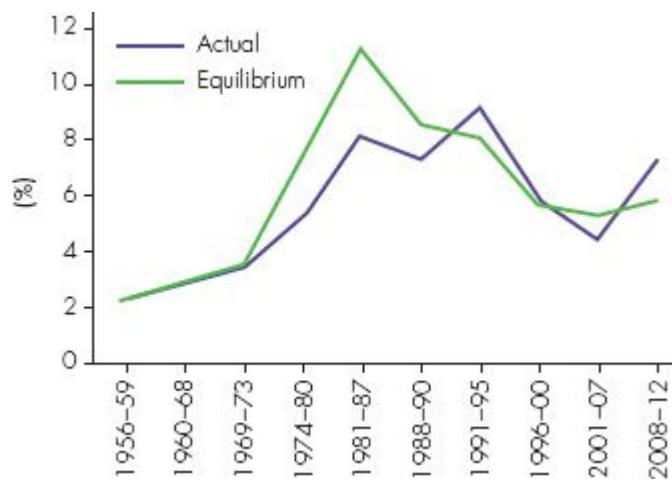


Figure 23.6 UK unemployment, 1956–2012 (% annual average)

Source: R. Layard, S. Nickell and R. Jackman, *Unemployment* (Oxford University Press, 1991); S. Nickell, 'Inflation and the UK Labour Market', in T. Jenkinson (ed.), *Readings in Macroeconomics* (Oxford University Press 1996); authors' estimates.

Until the 1970s, demand management maintained aggregate demand in line with aggregate supply in the output market. Hence, in the labour market, actual and equilibrium unemployment pretty much coincided. When unemployment then rose during the 1970s, people initially assumed that this must be due to deficient demand.

In retrospect, we know that the diagnosis was wrong – it was really the supply side that was deteriorating. Equilibrium unemployment was increasing steadily as work incentives deteriorated and the economy was hit with adverse supply shocks. Misreading the situation, governments tried boosting demand to eliminate spare capacity. Since they had no spare capacity, instead they stoked up inflation.

When Mrs Thatcher came to power in 1979 the Conservative government reduced aggregate demand to try to tackle inflation, and embarked on supply-side reform to reduce equilibrium unemployment. Figure 23.6 shows that demand-deficient unemployment rose sharply when aggregate demand was first reduced, whereas it took longer to obtain reductions in equilibrium unemployment.

Nevertheless, the rise in equilibrium unemployment was slowly reversed. The government became overconfident in its success, allowing aggregate demand to increase sharply during the ‘Lawson boom’ of the late 1980s. Figure 23.6 shows that actual unemployment had fallen below equilibrium unemployment as the economy overheated and inflation picked up again.

The 1990s saw the restoration of balance between demand and supply, and the continuing benefits of supply-side reform. By the late 1990s the UK was enjoying a period of low inflation, low unemployment and considerable stability. Figure 23.6 shows that actual and equilibrium unemployment were close together until the financial crash, after which aggregate demand plummeted and a new gap emerged between actual and equilibrium unemployment.

Figure 23.6 confirms three periods in which involuntary unemployment became important – the Thatcher squeeze in the early 1980s, the Major squeeze in the early 1990s and the aftermath of the financial crash of 2008. In the first two examples, sharp reductions in demand were the result of policies to combat inflation.

However, the main message of Figure 23.6 is that longer-term trends in unemployment have been largely caused by supply-side changes and

their consequences for equilibrium unemployment.

Supply-side factors

Keynesians believe that the economy can deviate from full employment for quite a long time, certainly for several years. Monetarists believe that the classical full-employment model is relevant much more quickly. Everyone agrees that long-run performance is changed only by affecting the level of full employment and the corresponding level of potential output.

We now discuss four reasons why equilibrium unemployment rose and then fell during 1970–2010.

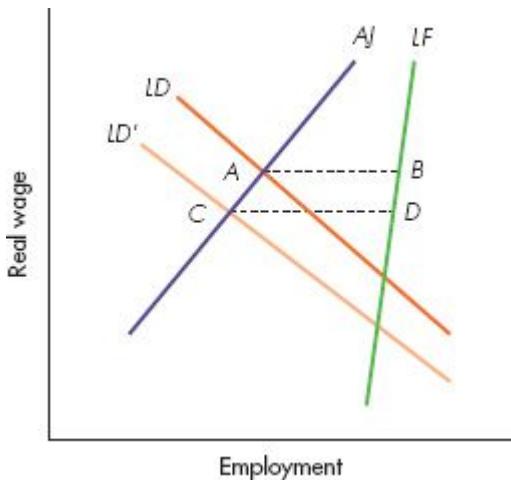
First, increasing skill **mismatch** raised equilibrium unemployment after 1970. Recent research emphasizes that the labour market is not very good at processing workers as they step out of one job and hope to step into another. The larger is mismatch, the harder the task is to perform, and the more likely it is that people get stuck in unemployment.

Supply-side economics is the use of microeconomic incentives to alter the level of full employment, the level of potential output and equilibrium unemployment.

Mismatch occurs if the skills that firms demand differ from the skills the labour force possesses.

When firms no longer want the skills possessed by the existing workforce, the labour demand curve LD shifts leftwards to LD' in Figure 23.7, leading to a lower equilibrium real wage, and an increase in equilibrium unemployment from AB to CD . A rise in mismatch explained some of the rise in unemployment in the 1970s and 1980s.

Conversely, since 1990, government policy has stressed reconnecting the unemployed with the labour market rather than leaving them to languish in long-term unemployment. By offering the unemployed advice on how to get back into work quickly, government policy stopped people becoming stigmatized as unemployable. This raised the demand for their labour, reducing equilibrium unemployment. At a higher real wage, AJ and LF are closer together.

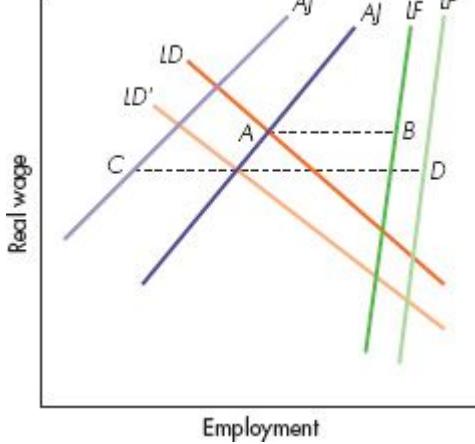


When labour demand is LD , equilibrium unemployment is AB . If mismatch reduces the demand for the skills that the workers possess, labour demand shifts down to LD' and equilibrium unemployment increases to CD .

Figure 23.7 An increase in mismatch

A second potential explanation of a rise in equilibrium unemployment is a rise in the generosity of unemployment benefit relative to wages in work. A higher **replacement rate** may entice more people into the labour force, shifting LF to the right. More significantly, it shifts AJ to the left. People spend longer in unemployment searching for the right job. For both reasons, equilibrium unemployment increases in Figure 23.8.

The **replacement rate** is the level of benefits relative to wages in work.



A higher replacement rate shifts the labour force from LF to LF' , and job acceptances from AJ to AJ' , raising equilibrium unemployment from AB to CD .

Figure 23.8 A higher replacement rate

Most empirical research concludes that higher benefits caused some of the increase in equilibrium unemployment, though less than sometimes supposed. In practice, UK unemployment benefit (now jobseeker's allowance) did not rise enough to explain the rise in unemployment.

However, benefits policy probably does explain some of the fall in equilibrium unemployment after 1992. First, as in other countries such as the Netherlands, the UK redefined many of its long-term unemployed as sick. People on sickness benefit are no longer measured as unemployed. This improves statistical unemployment, though of course in economic terms it is entirely cosmetic.

Second, Labour's employment policy viewed getting the unemployed back into work as the best form of social policy. People reacquire the work habit and rebuild their confidence. Accordingly, Labour focused on its *Welfare to Work* and *Making Work Pay* – measures intended, respectively, to actively assist the unemployed to look for work and to incentivize them to want to look for work. Measures of these types are sometimes called *active labour market policy*.

Since 2010 the coalition government has begun another reform of welfare policy designed to enhance work incentives further and/or to reduce the fiscal cost of benefits (and thereby the tax rate levied on other workers in order to keep the Treasury solvent). New measures include capping the rate of increase of nominal benefit levels, in effect ensuring a gradual reduction of the real benefit level for those out of work. We discuss these in more detail in Activity 23.1

A third source of changes in equilibrium unemployment has been changes in **trade union power**.

Trade union power is measured by the ability of unions to co-ordinate lower job acceptances, thereby increasing wages but reducing employment.

Rises in union power, especially in the 1970s, had a big effect on equilibrium unemployment. Powerful unions made labour scarce and

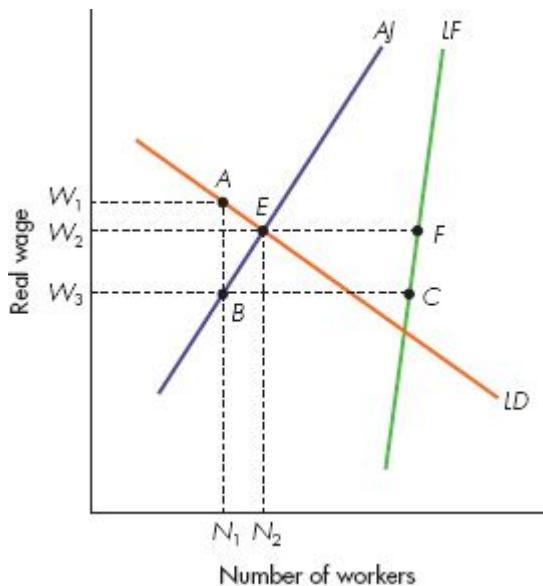
forced up its price. By shifting the *AJ* curve to the left, unions forced up real wages but increased equilibrium unemployment. Conversely, the fall in union power has shifted the *AJ* schedule right, reducing equilibrium unemployment.

Union power increased in the 1970s partly because sympathetic governments passed legislation enhancing worker protection and partly because many nationalized industries were sheltered state monopolies from which unions could extract potential profits as extra wages for their members. Their power declined after the 1980s, partly because a less sympathetic government reduced the legal protection of unions, privatization removed the Treasury as last-resort funder of union wage claims and globalization increased competition in general. The combination of (a) privatization, (b) globalization, (c) the advent of new technologies, particularly the Internet, which enable more flexible working, and (d) the continuing rise of the service sector and decline of large manufacturing industry has considerably reduced the significance of trade unions over the last 40 years.

The final important source of changes in equilibrium unemployment was changes in the size of the **tax wedge** between the cost of labour to the firm and the take-home pay of the worker. A key theme of supply-side economists is the benefits that stem from reducing the **marginal tax rate**.

The **marginal tax rate** is the fraction of each extra pound that the government takes in tax. This creates a **tax wedge** between the price the purchaser pays and the price the seller receives.

A cut in marginal tax rates, and a consequent increase in the take-home pay derived from the last hour's work, make people substitute work for leisure. Against this *substitution effect* must be set an *income effect*. If people pay less in taxes, they have to do less work to reach any given living standard target. Thus, theoretical economics cannot prove that tax cuts raise desired labour supply. Most empirical studies confirm that, at best, tax cuts lead to only a small rise in labour supply. Figure 23.9 shows how tax rates affect equilibrium unemployment.



An income tax makes the net-of-tax wage received by households lower than the gross wage paid by firms. AB measures the amount each worker pays in income tax, and equilibrium employment is N_1 , the quantity that households wish to supply at the after-tax wage W_3 and that firms demand at the gross wage W_1 . At the after-tax wage W_2 the natural rate equilibrium would be at EF . Employment would rise from N_1 to N_2 and the natural rate of unemployment would fall from BC to EF . Relative to the fixed level of unemployment benefit, the rise in take-home pay from W_3 to W_2 reduces voluntary unemployment.

Figure 23.9 A cut in marginal income tax rates

Suppose the marginal tax rate equals the vertical distance AB . Equilibrium employment is then N_1 . The tax drives a wedge between the gross-of-tax wages paid by firms and the net-of-tax wages received by workers. Firms wish to hire N_1 workers at the gross wage w_1 . Subtracting the income tax rate AB , N_1 workers want to take job offers at the after-tax wage w_3 . Thus N_1 is equilibrium employment, where quantities supplied and demanded are equal. The horizontal distance BC shows equilibrium unemployment: the number of workers in the labour force not wishing to work at the going rate of take-home pay.

Suppose taxes are abolished. The gross wage and the take-home pay now coincide, and the new labour market equilibrium is at E . Two things happen. First, equilibrium employment rises. Second, although more people join the labour force because take-home pay has risen from w_3 to w_2 , equilibrium unemployment falls from BC to EF . A rise in take-home

pay relative to unemployment benefit reduces voluntary unemployment. If lower tax rates reduce equilibrium unemployment, higher tax rates increase equilibrium unemployment.

ACTIVITY 23.1

DOES THE TAX CARROT WORK?

A lower marginal tax rate on income makes people substitute work for leisure. Since working is now better rewarded, people choose to work a bit more. But tax cuts also make workers better off by raising disposable income earned from any particular number of hours of work. This income effect makes people want to consume more leisure, and hence work fewer hours. The combined effect on hours of work is small for those already in work. The income and substitution effects roughly cancel out.

Of more importance is the decision about whether to work at all. Higher take-home pay, for example because of tax cuts, makes more people join the labour force by reducing the significance of the fixed costs of working (commuting, finding cleaners and babysitters, giving up social security).

Over a long period, UK evidence has shown that tax cuts had a tiny effect on labour supply by the primary breadwinner in a household. But in households where both partners consider working, higher take-home pay encourages labour force participation by encouraging the second partner to work, overcoming the fixed costs of going out to work.

The Thatcher programme

The most dramatic natural experiment with which to assess the effectiveness of tax cuts is the programme of the Thatcher government in the 1980s. The real value of personal allowances – how much you can earn before paying income tax – rose by 25 per cent. The basic rate of income tax fell from 33 to 22 per cent and, for top income-earners, from 83 to 40 per cent. Many politicians anticipated a surge in labour supply. Most economists were pessimistic because of the evidence from the past.

The effect of the Thatcher programme was assessed by C. V. Brown in 1988.¹ Brown found that the big rise in tax allowances led to less

than 0.5 per cent extra hours of labour supply. The cut in the basic rate of income tax had no detectable effect at all. The massive cut in the marginal tax rate of top earners had a small effect in stimulating extra hours of work by the rich. The evidence from the past stood up well to a big change in tax policy.

New Labour after 1997

During 1997–2001 Chancellor of the Exchequer Gordon Brown quietly raised taxes to help the poor and provide funding for public services. In order not to scare the middle classes, the government kept rather quiet about these tax increases, which were sometimes labelled ‘stealth taxes’. Neither theory nor past evidence suggests that these had a large and adverse incentive effect. We know from Figure 23.6 that equilibrium unemployment remained low thereafter.

It remains to be seen whether some sharper adjustments required by the labour market as the economy reacts to the financial crash and its austerity aftermath will lead to a renewed rise in equilibrium unemployment; for example, because the composition of demand in the new economy requires substantially different skills, as would be the case if the UK experienced another shakeout of old manufacturing jobs and the further rise of new service industries. We can be fairly sure that there will be a period of demand-deficient unemployment.

Emigrating non-doms and hedge fund managers

A general lesson of tax policy is that, when things are very elastically supplied or demanded (very price sensitive), tax rates can have large effects on the quantity traded. Conversely, when things are very inelastic in supply or demand (very price insensitive), tax rates have little effect on the quantity traded.

Since most workers are reluctant either to emigrate or to give up working, income tax usually has only a small effect on labour supply. But there are exceptions. Investment bankers, hedge fund managers and the super-rich may sometimes fall into this category. If they have global lifestyles anyway, they may be relatively indifferent as to whether they live in Manhattan, London, Paris or Geneva. If a single country attempts to tax them very differently

from other countries, they may move location in order to find a more agreeable tax regime.

Of course, they are likely to threaten to move, whether or not they really will. This may deter a government from trying to tax them more heavily. In 2009 the UK announced both that it was going to raise its top income tax rate from 40 to 50 per cent in 2010, and also proposed eliminating some of the tax breaks enjoyed by rich foreign residents (the so-called non-domiciles or non-doms). In the subsequent outcry, the government backed down on non-doms, but the top marginal income tax rate was increased to 50 per cent as planned.

The Coalition Reforms

In 2013, the coalition government cut this top rate from 50 to 45 per cent in order to mitigate perverse incentive effects. This will provoke a new burst of empirical research to see if any beneficial effects can be detected. If the evidence from the past continues to hold up, such effects will be disappointingly small.

Finally, it should be remembered that it is the combined effect of the tax and benefit systems that determines work incentives. For poorer people, the benefits system is much more important than the tax system. Under New Labour (1997–2010) there was a focus on providing in-work benefits to try to get as many people as possible into the labour force. Given the modern prevalence of part-time working, New Labour decided to make these benefits available only for people working at least 16 hours a week (reckoned to be what a single mother with children could manage). Hence quite a large number of part-time workers clustered around 16 hours a week of part-time work, thereby getting additional welfare support.

Since 2010, UK government policy on welfare has focused on the possibility of a ‘universal benefit’ that everyone gets, combined with steady loss of benefits as income rises. Effectively, this raises the marginal tax rate of the ‘tax plus benefit’ system in the income range in which people are losing benefit by working longer.

This system is ‘less distortionary’ than the previous one, since it provides continuous and smooth incentives rather than big discontinuities at particular levels of work. In principle, that is a good idea. Critics argue that it may nevertheless be

counterproductive if its main effect is to remove the incentive for potential part-time workers to work a minimum of 16 hours. Under the new system, they may choose to work less than that.

Questions

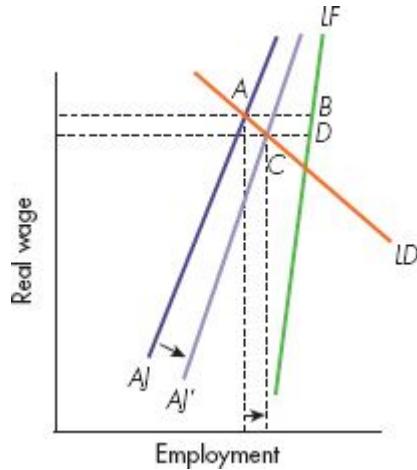
- a. Suppose national insurance contributions by firms – which depend on the value of wages paid to workers – are increased by 1 per cent. Draw a diagram to display the effect on equilibrium unemployment. Does this constitute a ‘tax on jobs’?
- b. Now suppose that, instead of higher national insurance contributions by firms, the same additional revenue for the Treasury is raised by asking workers to increase the contributions they make to national insurance. How, if at all, does the outcome differ from that in (a)?
- c. Suppose that the revenue raised is spent entirely on additional nurses for the National Health Service. Is the combined effect of the two policies a tax on jobs?

To check your answers to these questions, go to page 681.

Another possible supply-side policy is to cut unemployment benefit. For a given labour force schedule LF , fewer people now wish to be unemployed at any real wage. In Figure 23.10, the schedule AJ , showing acceptances of job offers, shifts to the right. This raises equilibrium employment (and hence potential output) and reduces equilibrium unemployment.

What about changes in the national insurance contributions paid both by firms and by workers? These are mandatory contributions to state schemes that provide unemployment and health insurance. They act like an income tax, driving a wedge AB between the total cost to a firm of hiring another worker and the net take-home pay of a worker. Figure 23.9 implies that a fall in these contributions will raise equilibrium employment and cut equilibrium unemployment.

Supply-side policies can reduce equilibrium unemployment. Where this involves being tough on those already relatively disadvantaged, there is a conflict between efficiency and fairness, and only through the political process can society express its view.



Initially, the job acceptance schedule is AJ and equilibrium unemployment is the distance AB . A fall in unemployment benefit raises the cost of searching for a better job offer. Job acceptances shift to AJ' and equilibrium unemployment falls to the distance CD .

Figure 23.10 A reduction in unemployment benefit

MATHS 23.1

THE LABOUR MARKET INCIDENCE OF POLICY CHANGES

Suppose the labour force L , job acceptances A and labour demand D obey the equations:

$$L = a + bw \quad A = (b + c)w \quad D = e - f(w + t) \quad a, b, c, e, f > 0, t > 0 \quad (1)$$

where a, b, c, e and f are positive constants, and t is the effective tax rate on working. Workers care about take-home pay w , whereas firms care about the gross cost $(w + t)$ of hiring a worker. In labour market equilibrium, job acceptance A equals labour demand D , whence

$$w = (e - ft)/(b + c + f) \quad A = (e - ft)(b + c)/(b + c + f) \quad (2)$$

and equilibrium unemployment ($L - A$), which is simply $(a - cw)$, is given by

$$u^* = \{a(b + c + f) - c(e - ft)\}/(b + c + f) \quad (3)$$

Does a rise in labour taxes t increase equilibrium unemployment? In the model, the answer is yes – a one-unit increase in t increases u^* by $cf/(b + c + f)$. In terms of figures such as 23.7 and 23.8, once we assume that job acceptances are more sensitive than the labour force to increases in take-home pay, anything that reduces take-home pay widens the gap between those wishing to work and those accepting jobs. Equation (2) above confirms that higher labour taxes reduce the number of job acceptances by reducing the component $(e - ft)$ in take home pay. Check for yourself that labour demand D is reduced by the same amount. This occurs because the gross cost of labour ($w + t$) increases despite the fall in w , as implied by equation (2):

$$w + t = [e + (b + c)t]/[b + c + f]$$

which unambiguously increases with the tax rate t . In policy terms, the key issue is how much more sensitive to wage increases is the job acceptance schedule in comparison with the labour force schedule. If the two schedules had equal slope ($c = 0$), then equation (3) implies that the tax rate would have *no effect* on equilibrium unemployment. Intuitively, in terms of the diagrams earlier in the chapter, when the LF and AJ schedules are parallel, then the changes in the equilibrium quantity of employment, for whatever reason, have no effect on the horizontal gap that corresponds to the equilibrium rate of unemployment. However, the more responsive is the job acceptance schedule (larger c , flatter slope), the more a reduction in equilibrium employment widens the gap between job acceptances and the labour force, raising the equilibrium rate of unemployment.

23.4 Cyclical fluctuations in unemployment

We discuss business cycles in Chapter 27. Cycles may reflect fluctuations in demand or fluctuations in supply. Since supply usually changes slowly, most of the sharp movements in the short run are caused by changes in demand.

Unless a counter-cyclical demand management policy is deliberately and successfully pursued, there may well be a business cycle. If so, there tends to be a cyclical relationship between demand, output, employment and unemployment. On average, boosting aggregate demand by 1 per cent will not raise employment by 1 per cent or reduce unemployment by 1 per cent, even if the economy begins with spare resources. Table 23.3 shows two periods of demand growth and two of demand decline. In practice, booms lead initially to a sharp increase in shift lengths and hours worked; slumps lead to the abolition of overtime, the introduction of short time and a marked decline in hours worked.

Table 23.3 Output, employment and unemployment: evidence from the past

Cumulative change in	79ii–81ii	86ii–88ii	90ii–91ii	92iv–98ii
Real GDP (%)	-7.8	+9.1	-3.4	+16.8
Employment (%)	-6.3	+2.5	-2.9	+6.8
Employed (million)	-1.7	+0.5	-0.7	+1.5
Unemployed (million)	+1.4	-0.9	+0.6	-1.2

Source: ONS, *Economic Trends*.

The table confirms that changes in demand and output lead to smaller changes in employment. For example, when output grew by 16.8 per cent between the fourth quarter of 1992 and the second quarter of 1998, employment rose by only 6.8 per cent. Nor do changes in employment lead to corresponding changes in unemployment. The last two rows of the table show that rapid expansion or contraction of employment leads to significantly smaller changes in unemployment.

One reason is the ‘discouraged worker effect’. When unemployment is high and rising, some people who would like to work become pessimistic and stop looking for work. No longer registered as looking for work, they are not recorded in the labour force or the unemployed. Conversely, in a boom, people who had previously given up looking for work rejoin the labour force since there is now a good chance of getting a suitable job. Hence in booms and slumps recorded employment data change by more than recorded unemployment data. After 1997, the Monetary Policy Committee kept the UK economy on a more even keel until the recession of 2009.

CONCEPT 23.2

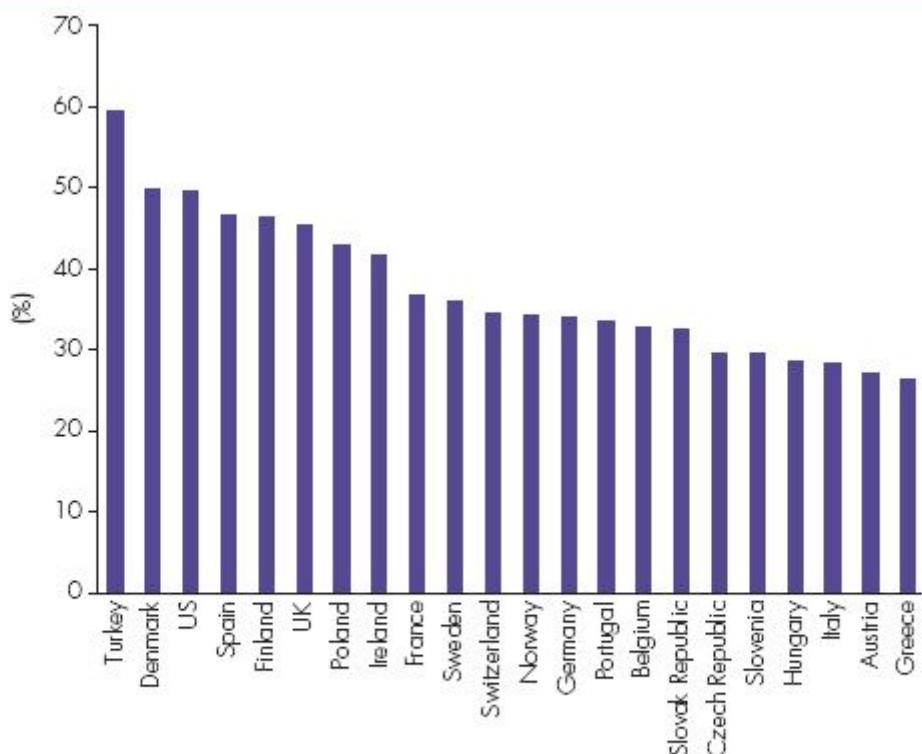
HOW BAD COULD UNEMPLOYMENT BECOME?

The OECD is a club of the most advanced economic nations in the world – living standards and per capita income count more than absolute size. Newer members include Turkey, Mexico and Hungary. China and India are not yet members despite their vast populations. Currently, the OECD has 32 members.

The 2009 OECD *Employment Outlook* discusses prospects for unemployment and possible policy responses. The evolution of unemployment depends on: (a) the size of the shock, (b) the flexibility of the economy to respond, and (c) the extent of support by government. In the worst previous post-war recession, that of 1973–76, OECD unemployment increased by half. By 2009 the OECD reckoned that unemployment would rise by 80 per cent – from 5.5 to 10 per cent of the labour force – during 2007–10.

This analysis reflected the magnitude of the initial shock. Clearly, this would affect different countries differently. One way in which to assess which economies were most vulnerable is to estimate their capacity to absorb shocks through flexible labour markets that match potential workers and job opportunities more quickly. This is likely to depend on wage flexibility, labour market mobility, attitudes of trade unions and the extent of labour market regulation.

The figure below shows a measure of labour market flexibility based on labour market history during 2000–05. It plots the annual fraction of workers hired in new jobs or leaving existing jobs (by choice or dismissal) during the year. It shows that, in Turkey, Denmark and the US, half of all workers are changing jobs annually. In contrast, the countries with the lowest labour market mobility are Greece, Italy and Austria.



Annual fraction of workers hired or fired, 2000–05 (%)

Source: OECD, *Employment Outlook, 2009*.

Countries with greater job stability are probably slower to experience initial unemployment but, when unemployment does increase, they are also less successful at helping people out of unemployment back into work. Since there is considerable cross-country evidence that those in longer-term unemployment find it ever more difficult to reconnect with the labour market, in the medium run this fiscal burden of unemployment benefits is likely to be greater in countries with less flexible labour markets.

Duration of unemployment	Year 1	Year 2	Year 5
	Replacement rate (%)		
Norway	72	72	72
Belgium	65	63	63
France	67	64	31
UK	28	28	28
Japan	45	3	3
Greece	33	5	1
US	28	0	0

Source: OECD, Employment Outlook, 2009.

Governments provide two kinds of support. The first is measurable by the generosity of unemployment benefit, which has two dimensions – the replacement rate (the ratio of benefit to previous wages in work) and the number of years for which benefit is available. The table documents considerable differences across countries.

In Norway and Belgium, with strong traditions of social democracy, unemployment benefit is generous both because it is high relative to wages in work and because it continues for at least five years after a spell of unemployment begins. French unemployment benefit is initially as generous but less so after year two. The UK is considerably less generous in its replacement rate, but entitlement continues undiminished over the five-year period. In countries such as Japan, Greece and the US, unemployment benefit is almost worthless after the first year in unemployment.

The second aspect of state support for the unemployed is active labour market policies that enhance incentives, confidence and the ability of the unemployed to look for jobs. Even if the post-crash recession reflected a sharp fall in demand – for output and then for labour – it is important not to neglect supply-side policies that maintain maximum labour market flexibility.

With the benefit of hindsight, the most puzzling features of the labour market since the financial crash have been (a) the relatively small increase in unemployment and (b) the disappointing performance of labour productivity. For example, UK unemployment was still only 8 per cent in late 2012. The US has also avoided double-digit unemployment and appeared to have resumed steady if modest output growth. Two explanations consistent with facts (a) and (b) are that (i) firms engaged in more labour hoarding in recession than in previous cyclical recessions, and (ii) that many workers who would otherwise have become unemployed took up part-time working or self-employment. The rise of the Internet allowed many people to embark on small businesses in the service sector.

With long-term demographic trends also implying that pension schemes would face greater and greater financial strains, and with governments removing subsidies to pension contributions in an

effort to retain fiscal solvency, it is also possible that some older workers simply retired while the going was good, rather than face a period of unemployment when rehiring of older workers was likely to be a tough prospect.

23.5

The cost of unemployment

The private cost of unemployment

It is important to distinguish between voluntary and involuntary unemployment. When individuals are voluntarily unemployed, they reveal that they do better by being unemployed than by immediately taking a job offer at the going wage rate. The private cost of unemployment (the wage forgone by not working) is less than the private benefits of being unemployed. What are these benefits?

The first is transfer payments from government. Workers who have contributed to the national insurance scheme get jobseeker's allowance for the first 12 months after becoming unemployed. Thereafter they get income support, the ultimate backstop in the British welfare state.

There are other benefits, too. First, there is the value of leisure. By refusing a job, some people reveal that the extra leisure is worth more to them than the extra disposable income if they took a job. Second, some people expect to get a better job by being choosy about accepting offers. These future benefits must be set against the current cost: a lower disposable income as a result of being out of work.

When people are involuntarily unemployed, the cost changes. Involuntary unemployment means that people would like to work at the going wage but cannot find a job because there is excess labour supply at the existing wage rate. These people are worse off by being unemployed.

The distinction between voluntary and involuntary unemployment matters because it may affect our value judgement about how much attention to pay to unemployment. When unemployment is involuntary, people are suffering more and the case for helping them is stronger.

The social cost of unemployment

Again we distinguish between voluntary and involuntary unemployment. When unemployment is voluntary, individuals prefer to be unemployed. Does this unemployment also benefit society?

An individual receives transfer payments during unemployment, but these transfers give no corresponding benefit to society as a whole. They may ease the collective conscience in regard to poverty and income inequality, but they are not payments for the supply of any goods or services that other members of society may consume. Since the private benefit exceeds the social benefit, too many people may be voluntarily unemployed.

CONCEPT 23.2

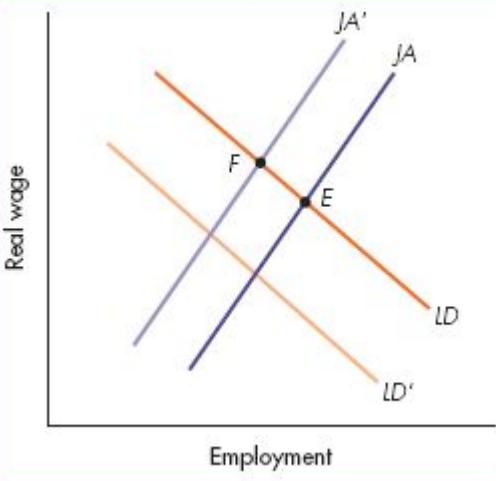
HYSTERESIS AND HIGH UNEMPLOYMENT

Supply and demand curves are supposed to be independent of one another. The labour supply curve or job acceptances schedule AJ shows the people willing to work at each real wage whatever the position of the labour demand curve LD , and vice versa. But this may be wrong.

In the diagram, the initial equilibrium is at E . Something then shifts labour demand down from LD to LD' . Suppose this causes a permanent fall in labour supply. JA shifts to JA' . When labour demand reverts to LD , the new equilibrium is at F , not E . The short-run history of the economy has affected its long-run equilibrium.

Hysteresis may explain high and persistent unemployment in much of continental Europe. Here are some channels through which it might work.

An economy experiences **hysteresis** when its long-run equilibrium depends on the path it follows in the short run.



The insider–outsider distinction

Outsiders are those unemployed without jobs. Only insiders with jobs participate in wage bargaining. At the original equilibrium E , the numerous insiders in work ensure that real wages are low enough to preserve their own jobs. When a recession occurs, LD shifts to LD' . Some insiders get fired and become outsiders. Eventually, as explained in Chapter 21, market forces restore labour demand to LD . But now there are fewer insiders than originally. They exploit their scarcity by securing higher wages for themselves rather than encouraging firms to rehire. The economy is trapped in the high-wage, low-employment equilibrium at F instead of the low-wage, high-employment equilibrium at E . Thereafter, only long-run supply-side measures aimed at breaking down insider power can gradually break the economy out of this low-employment equilibrium.

Discouraged workers

Again, the economy begins at E . It has a skilled and energetic labour force. A temporary recession leads to unemployment. If the recession is protracted, we see the emergence of long-term unemployed people and a culture in which they stop looking for jobs. Again, when demand picks up, labour supply has been permanently reduced and equilibrium reverts to F , not E . Only long-term supply-side measures to restore the work culture will succeed.

Search and mismatch

When employment is high at E , firms are trying to find scarce workers, and potential workers are searching hard for a job. A recession makes firms advertise fewer vacancies, and workers realize it is a waste of time searching for jobs. When demand picks up again, both firms and workers are accustomed to low levels of search. New jobs are not created.

The capital stock

At E , the economy has a lot of capital. Labour productivity is high and firms want lots of workers. During a temporary recession, firms scrap old machines. When demand picks up again, firms have permanently lower capital. The demand for labour, which depends on the marginal product of labour, never rises to its original level. Again, the economy returns to F , not E .

Policy implications of hysteresis

Hysteresis means that a temporary fall in demand induces permanently lower employment and output, and higher equilibrium unemployment. There are two policy implications. First, once the problem has emerged, it is dangerous to try to break out of it simply by expanding aggregate demand. Before long-run supply can respond, you get major inflation. Supply-side policies, needed to rebuild aggregate supply, take a long time to work.

Second, because the problem is so hard to cure once it occurs, it is vital not to let demand fall in the first place. The payoff to demand management is higher than in an economy with a unique long-run equilibrium where all that is at stake is how quickly the economy reverts to its original point.

These arguments help explain why governments intervened so heavily in 2009–10 to endeavour to offset the worst of the demand effects of the financial crash. They feared that too little action would imply a dramatic fall in aggregate demand from which it would be hard to recover.

Empirical evidence

A useful survey of the evidence that hysteresis exists is provided by Laurence Ball.² Examining data for 20 countries, Ball found that large increases in *equilibrium* unemployment occur after a period of disinflation, as the figure above would imply.

This does not mean that society should go to the opposite extreme and eliminate voluntary unemployment completely. First, society is perfectly entitled to adopt the value judgement that it will maintain a reasonable living standard for the unemployed, whatever the cost in resource misallocation. Second, the efficient level of voluntary unemployment is well above zero.

In a changing economy, it is important to match up the right people to the right jobs. Getting this match right lets society make more output. Freezing the existing pattern of employment in a changing economy leads to a mismatch of people and jobs. The flow through the pool of unemployment allows people to be reallocated to more suitable jobs, raising potential output in the long run.

Two points from our earlier discussion are also relevant here. First, even when unemployment is high, flows both into and out of the pool are large relative to the pool itself. Second, people who do not get out of the pool quickly are in danger of stagnating when unemployment is high: the fraction of the unemployed who have been unemployed for over a year was higher in the 1990s than at the end of the 1970s when unemployment was much lower.

Involuntary or Keynesian unemployment has an even higher social cost. Since the economy is producing below capacity, it is literally throwing away output that could have been made by putting these people to work. Moreover, since Keynesian unemployment is involuntary, it may entail more human and psychological suffering than voluntary unemployment. Although hard to quantify, it is also part of the social cost of unemployment.

Summary

- People are either **employed, unemployed** or out of the **labour force**.
The level of unemployment rises when inflows to the pool of the unemployed exceed outflows. Inflows and outflows are large relative to the level of unemployment
- As unemployment has risen, the average duration of unemployment has increased.

- Women face lower unemployment rates than men. The unemployment rates for old workers and, especially, for young workers are well above the national average.
- **Unemployment** can be classified as **frictional, structural, classical** or **demand-deficient**. In modern terminology, the first three types are **voluntary unemployment** and the last is **involuntary unemployment**. The **natural rate of unemployment** is the equilibrium level of voluntary unemployment.
- In the long run, sustained rises in unemployment must reflect increases in the natural rate of unemployment. During temporary recessions, **Keynesian unemployment** is also important.
- **Supply-side economics** aims to increase equilibrium employment and potential output, and to reduce the natural rate of unemployment, by operating on incentives at a microeconomic level. Supply-side policies include reducing mismatch, reducing union power, tax cuts, reductions in unemployment benefit, retraining and relocation grants and investment subsidies.
- A 1 per cent increase in output is likely to lead to a much smaller reduction in Keynesian unemployment. Some of the extra output will be met by longer hours. And as unemployment falls, some people, effectively in the labour force but not registered, look for work again.
- **Hysteresis** means that short-run changes can move the economy to a different long-run equilibrium. It may explain why European recessions have raised the natural rate of unemployment substantially.
- People voluntarily unemployed reveal that the private benefits from unemployment exceed the private cost in wages forgone. Society derives no output from transfer payments to support the unemployed. However, society would not benefit by driving unemployment to zero. Some social gains in higher productivity are derived from improved matching of people and jobs that temporary unemployment allows.

- Keynesian unemployment is involuntary and hurts private individuals who would prefer to be employed. Socially it represents wasted output. Society may also care about the human misery inflicted by involuntary unemployment.
- Most European countries took two decades to reverse the high unemployment of the 1980s. Having brought unemployment down, they now face a new period of higher unemployment as budgets are cut back to cope with the fiscal crises caused by trying to offset the effects of the financial crash.

Review questions



EASY

- 1 How is high unemployment explained by (a) a Keynesian and (b) a classical economist?
- 2 Which of the following statements is correct? The fact that unemployment rose in 2009 in the UK by less than originally predicted shows that: (a) the fall in output and demand was illusory; (b) wages were more flexible than in previous recessions; (c) firms believed that lower output would be very short lived?
- 3 What is the discouraged worker effect? Suggest two reasons why it occurs.
- 4 The average duration of an individual's unemployment rises in a slump. Hence the problem is a higher inflow to the pool of unemployment, not a lower outflow.' Do you agree?
- 5 Explain why boosting demand sometimes fails to reduce unemployment.

MEDIUM

- 6 'The microchip caused a permanent rise in the level of unemployment.' Did it? What about all previous technical advances?
- 7 Illustrate on a graph the effect of a labour skill mismatch on the equilibrium level of unemployment in the labour market.
- 8 'An increase in national insurance contributions by workers reduces the income per hour that workers take home and therefore reduces the

incentive to work.' 'An increase in national insurance contributions, by reducing income per hour, forces people to work longer hours to attain their target take-home income.' Is either statement correct? Are both correct? What light does this shed on national insurance contributions as a 'jobs tax'? Draw a diagram to illustrate your answer.

- 9 Labour supply L , job acceptances J and labour demand D are, respectively, related to the real wage W by:

$$L = 10 + W, J = bW, 0 < b < 1 D = 50 - dW, d > 0$$

(a) Find equilibrium unemployment. (b) If there is now an income tax at rate t on wages, what happens to equilibrium unemployment?

- 10 Suppose the government wants to encourage lone parents to take part-time jobs and thinks 15 hours a week is consistent with children being in a crèche for 3 hours a day, Monday to Friday. Which of the following might achieve the government's aim: (a) an additional lump-sum payment to lone parents, (b) a lower income tax rate for lone parents, or (c) a payment conditional on their taking at least 15 hours of work a week?
- 11 **Common fallacies** Why are these statements wrong? (a) Unemployment is always a bad thing. (b) So long as there is unemployment, there is pressure on wages to fall. (c) Unemployment arises only because greedy workers are pricing themselves out of a job.

HARD

- 12 Why is teenage unemployment so high?

- 13 **Essay question** For two decades, unemployment in France has been significantly higher than that in the UK. If you become the president of France, should you: (a) blame the European Central Bank for cautious monetary policy; (b) blame the French Treasury for a fiscal policy that has been too tight; or (c) tackle labour market reform in France? Explain your answer.

- 14 Most economists forecast a period of protracted unemployment after 2009 as government takes tough measures for a sustained period to bring the budget deficit and national debt under control. (a) Why might such an evolution involve hysteresis? (b) Draw a diagram to illustrate both the initial increase in unemployment and the subsequent developments when demand eventually expands again.

-
- 1 C. V. Brown, ‘The 1988 tax cuts, work incentives and revenue’, *Fiscal Studies* 9, no. 4 (1988): 93–107.
 - 2 Laurence Ball, ‘Hysteresis in unemployment: Old and new evidence’, in J. Fuhrer (ed.), *A Phillips curve retrospective* (Federal Reserve Bank of Boston and MIT Press, 2009).

CHAPTER 24

Exchange rates and the balance of payments

Learning outcomes

By the end of this chapter, you should be able to:

- 1 analyse the foreign exchange market
- 2 discuss balance of payments accounts
- 3 explain determinants of current account flows
- 4 define perfect capital mobility
- 5 assess speculative behaviour and capital flows
- 6 define internal and external balance
- 7 analyse the long-run equilibrium real exchange rate

Exports and imports are each about 10 per cent of the size of GDP in Japan, 15 per cent in the US, around 30 per cent in the UK and France, 40 per cent in Germany, but over 70 per cent in small European economies such as Belgium. Even in the US and Japan, the exchange rate, international competitiveness and the trade deficit are major issues. International linkages in more **open** economies, such as the UK, Germany and Belgium, matter even more.

An **open economy** has important trade and financial links with other countries. A **closed economy** has no economic links with foreign countries.

In this chapter we show how international transactions affect the domestic economy.

24.1

The foreign exchange market

Different countries use different national currencies. In the UK, goods, services and assets are bought and sold for pounds sterling; in France, they are bought and sold for euros.

Measuring exchange rates

Suppose \$2 converts to £1. We can say either that the exchange rate is \$2/£ or that it is £0.50/\$. Both statements contain the same information.

Thus an exchange rate of \$2/£ is the **international value of the domestic currency** as viewed by a UK resident, but the domestic price of foreign exchange as viewed by a US resident. Conversely, £0.50/\$ is the **domestic price of foreign exchange** for a UK resident but the international value of the domestic currency for a US resident.

The **foreign exchange (forex) market** exchanges one national currency for another at a price called the **exchange rate**.

The **international value of the domestic currency** is the quantity of foreign currency per unit of the domestic currency.

The **domestic price of foreign exchange** is the quantity of domestic currency per unit of the foreign currency.

Whenever you see a table or graph with ‘the’ exchange rate, you need to work out which way round it has been expressed. There is no short cut. Even after years in the subject, we ourselves go slowly at that bit. For the rest of this book, we will use the international value of domestic currency. For the UK, this means talking about dollars or euros that exchange for £1. If we are discussing Germany, it would mean the number of pounds or dollars that exchange for 1 euro.

On television and at money-changing kiosks, you will rarely see quotes of £/\$ – most of the world conventionally quotes \$/£ whether they are talking about the US or the UK. This means that if you happen to look at a US textbook, you will find it using the domestic price of foreign

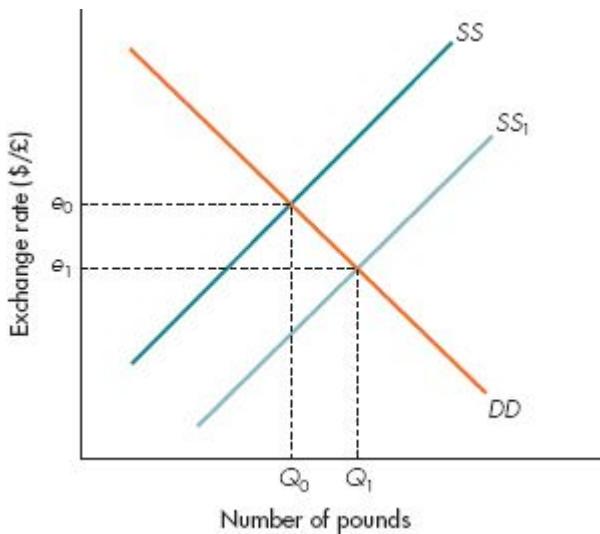
exchange as its definition of the exchange rate. When its graph slopes up and the text talks about its exchange rate ‘depreciating’, this is because it is thinking about the exchange rate the other way round from how someone in the UK would view it. A change from \$1/£ to \$2/£ is an appreciation of sterling but a depreciation of the dollar. Conversely, a change from \$2/£ to \$1/£ is a depreciation of sterling but an appreciation of the dollar.

For exchange rates against the euro, it is quite common to see them quoted both ways, either as €/£ or as £/€. Whatever the circumstances, take your time and ask which is the domestic currency of the country you are considering. Foreign currency per unit of domestic currency is the international value of the domestic currency. Domestic currency per unit of the foreign currency is the domestic price of foreign exchange.

Whichever way we express the exchange rate, in practice each currency exchanges for many others, not just one. However, for simplicity our discussion assumes only two countries, the domestic economy (say the UK) and the foreign country (say the US).

Exchanging currencies

Who supplies dollars to the forex market demanding pounds in exchange? The demand for pounds has two sources. First, US importers pay in dollars but UK exporters want to bring this money home as pounds. Second, US residents buying UK assets (shares in BT or UK bonds) must convert their dollars into pounds to buy these UK assets. Conversely, a supply of pounds reflects UK imports of US goods and UK residents buying assets in the US.



DD shows demand for pounds by Americans wanting to buy British goods or assets. SS shows the supply of pounds by British residents wishing to buy American goods or assets. The equilibrium exchange rate is e_0 . If British residents want more dollars at each exchange rate, the supply of pounds will shift from SS and the equilibrium international value of the pound will fall.

Figure 24.1 The forex market

Figure 24.1 shows the supply and demand for pounds in the forex market. We begin with the demand. Suppose UK whisky costs £8 a bottle. At \$2/£ it sells in the US for \$16, but at \$1.50/£ it sells for \$12. Hence at a lower exchange rate,¹ and a lower dollar price of all UK goods, the UK exports more goods to the US. US residents buy more at a lower dollar price.

If the sterling price of UK goods is constant, a lower exchange rate, by raising the quantity of UK exports, must raise export revenue in pounds. Figure 24.1 shows that the demand schedule for pounds, DD , slopes downwards. More pounds are demanded at a lower \$/£ exchange rate.

The supply of pounds SS depends on the quantity of dollars UK residents need to buy UK imports of goods or to buy dollar assets. Suppose a holiday in Florida costs \$600: at \$2/£ it costs £300, but at \$1.50/£ it costs £400. A lower \$/£ exchange rate raises the price in pounds and reduces the quantity of Florida holidays demanded by UK residents. Whether it reduces the number of pounds spent depends on the elasticity of demand for pounds.

Figure 24.1 assumes that the demand for Florida holidays and other UK imports is price-elastic. For a given dollar price of Florida holidays, a lower \$/£ exchange rate raises the price in pounds and reduces the sterling value of this spending. The supply of pounds *SS* slopes up. However, if the UK demand for US goods, services and assets is price-inelastic, a lower exchange rate and higher sterling price will raise sterling spending on these things, and the supply schedule of pounds to the forex market slopes down.²

At the equilibrium exchange rate e_0 , the quantity of pounds supplied and demanded is equal. What would change this equilibrium? Suppose, at each sterling price, US demand for UK goods or assets increases. The demand for pounds *DD* shifts to the right, raising the equilibrium \$/£ exchange rate. Similarly, a fall in UK demand for US goods and assets shifts the supply of pounds *SS* to the left, and the equilibrium \$/£ exchange rate rises.

When the \$/£ exchange rate rises, the pound **appreciates** so the dollar **depreciates**. Conversely, when the exchange rate is measured the other way, a rise in the £/\$ exchange rate reflects an appreciation of the dollar but a depreciation of the pound. This reinforces our earlier warning: to know whether a rise in the exchange rate reflects appreciation or depreciation, first you need to know which way round the exchange rate was measured.

Appreciation (of the exchange rate) is a rise in the international value of a currency.

Depreciation is a fall in the international value of a currency.

CONCEPT 24.1

EFFECTIVE EXCHANGE RATES

Each currency has a bilateral exchange rate against each other currency. For example, we can measure the \$/£ or €/£. Sometimes it is useful to examine a single exchange rate that summarizes all the bilateral rates.

The **effective exchange rate** (eer) is a weighted average of individual bilateral exchange rates. Usually, we use the share of

trade with each country to decide the weights. Important trading partners get more weight in the effective exchange rate index. The figure shows in purple sterling's effective exchange rate, based on an index whose value is set at 100 in June 2006. The figure also shows exchange rate indices for the UK's two main trading partners, the US and the Eurozone, both again set equal to 100 in June 2006. Sterling has fluctuated against both the dollar and the euro, but its average or effective rate is smoother than the individual exchange rates. If the euro and the US dollar were the *only* two currencies in the index, the purple line would *always* lie between the green and the orange lines. Since the euro and the dollar are much the most important currencies with which the UK trades, this is normally the case. But there are exceptions – for example, in 2012 the effective exchange rate lies below both the euro and the dollar indices. What are these other countries? There are lots – China, India, Brazil, all of Africa, the Middle East, Canada, Australia, and so on. But normally, what is happening with the US and the Eurozone is the bulk of the story.

A country's **effective exchange rate** is an average of its exchange rate against all its trade partners, weighted by the relative size of trade with each country.

The figure also implies that the weight on the euro is higher than the weight on the dollar: the effective or average exchange rate is more similar to the bilateral rate against the euro than to the bilateral rate against the dollar. Nowadays, the UK trades mainly with other European countries.

Once we have the nominal effective exchange rate, we could also construct a weighted average of foreign price levels and hence compute the real effective exchange rate. Changes in the real effective exchange rate are a simple indication of what is happening to competitiveness. Other things equal, higher real exchange rates imply lower competitiveness.



Sterling exchange rates, 1999–2012, (June 2006 = 100)

Source: Bank of England.

24.2 Exchange rate regimes

To grasp the basics, we focus on the two extreme forms of **exchange rate regime** that have been adopted to handle international transactions in the world economy: **fixed exchange rates** and floating exchange rates.

An **exchange rate regime** describes how governments allow exchange rates to be determined.

Fixed exchange rates

In Figure 24.2 suppose the exchange rate is fixed at e_1 . This is a free market equilibrium at A if the supply curve for pounds is SS and the demand curve for pounds is DD . Nobody needs to buy from or sell pounds to the central bank. The market clears unaided.

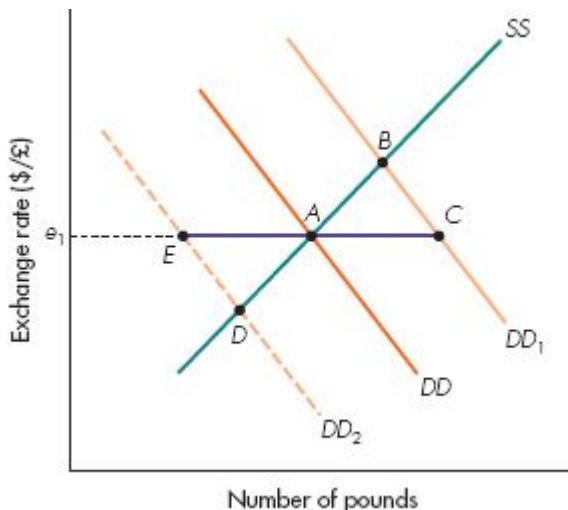
Suppose the demand for pounds shifts from DD to DD_1 . Americans, addicted to whisky, need more pounds to import more UK whisky. Free market equilibrium is now at B and the pound appreciates against the dollar. However, at a fixed exchange rate e_1 there is an excess demand for pounds equal to AC . To peg the exchange rate, the central bank meets this excess demand and maintains the peg e_1 by supplying an extra AC pounds to the market.

The Bank prints AC extra pounds and sells them in exchange for $(e_1 \times AC)$ dollars, which are added to the UK **foreign exchange reserves**.

In a **fixed exchange rate regime**, the currency is convertible at this exchange rate. A currency is *convertible* if the central bank will buy or sell as much of the currency as people wish to trade at the fixed exchange rate.

The **foreign exchange reserves** are foreign currency held by the domestic central bank.

What if the demand for pounds now falls to DD_2 ? The free market equilibrium is now at D . Pegging the exchange rate at e_1 causes an excess supply of pounds EA . To defend the peg the central bank must demand EA pounds, which it pays for by selling $(EA \times e_1)$ dollars from the foreign exchange reserves.



Suppose the exchange rate is fixed at e_1 . When demand for pounds is DD_1 , there is an excess demand AC . The Bank of England intervenes by supplying AC pounds in exchange for dollars, which are added to the UK foreign exchange reserves. When demand is DD_2 , the Bank sells foreign exchange reserves in exchange for pounds. It demands EA pounds to offset the excess supply EA . When demand is DD , the market clears at the exchange rate e_1 , and no intervention by the Bank is required.

Figure 24.2 Central bank intervention in the forex market

When the demand schedule is DD_1 , the UK is adding to its foreign exchange reserves. When the schedule is DD_2 , it is running down its reserves. If the demand for pounds fluctuates between DD_1 and DD_2 , the Bank of England can sustain the exchange rate e_1 in the long run.

However, if the demand for pounds on average is DD_2 , the Bank is steadily losing foreign exchange reserves to support the pound at e_1 . We say that the pound is overvalued, or is at a higher international value than is warranted by its long-run equilibrium position. As reserves start to run out, the government may try to borrow foreign exchange reserves from the International Monetary Fund (IMF), an international body that exists primarily to lend to countries in short-term difficulties.

At best, this is only a temporary solution. Unless the demand for pounds increases in the long run, it is necessary to *devalue* the pound. In a fixed exchange rate regime, a *devaluation (revaluation)* is a fall (rise) in the exchange rate governments commit themselves to maintain.

Notice that we say governments, plural. Fixing the \$/£ exchange rate is possible only if both the US and the UK wish to do so. For simplicity, our discussion of Figure 24.2 supposed that only one central bank intervened. In practice, it might be both central banks.

Floating exchange rates

In a **floating exchange rate regime**, the exchange rate is allowed to find its equilibrium level *without* central bank intervention using the forex reserves. Thus, in Figure 24.2, demand shifts from DD_2 to DD to DD_1 would be allowed to move the equilibrium from D to A to B .

In a **floating exchange rate regime**, the exchange rate is allowed to find its equilibrium level without central bank intervention using the

forex reserves.

Of course, it is not necessary to adopt the extreme regimes of pure or clean floating on the one hand and perfectly fixed exchange rates on the other. *Dirty floating* implies intervention is used to offset large and rapid shifts in supply or demand schedules in the short run, but the exchange rate is gradually allowed to find its equilibrium level in the longer run.

Having examined the foreign exchange market, we look next at the balance of payments.

24.3 The balance of payments

Taking the UK as the domestic country and the US as the ‘rest of the world’, all international transactions that give rise to an inflow of pounds to the UK are entered as credits in the UK **balance of payments** accounts. Outflows of pounds are debits, entered with a minus sign. Similarly, inflows of dollars to the US are credits in the US balance of payments accounts but outflows are debits. Table 24.1 shows the actual UK balance of payments accounts in the final quarter of 2011.

The **balance of payments** records transactions between residents of one country and the rest of the world. It is the sum of current, capital and financial account items.

Visible trade is exports and imports of goods (cars, food, steel). *Invisible trade* refers to exports and imports of services (banking, shipping, tourism). Together, these make up the trade balance or net exports of goods and services.

Current transfers are transfer payments paid across borders. These include payment by the UK government of EU subsidies for agriculture, social security payments paid abroad, bilateral foreign aid payments and cross-border flows of income, profits and dividends earned on assets held in other countries, less those earned on assets held by foreigners in the UK.

The **current account** of the balance of payments records international flows of goods, services and **current transfers** transfer payments paid

across borders.

Table 24.1 UK balance of payments, 2011 Q4 (£bn)

Trade in goods	-24
Trade in services	+17
Current transfers and other income	-2
(1) Current account	-9
(2) Capital account	+1
(3) Financial account	+16
(4) Balancing item	-8
(5) UK Balance of payments (1 + 2 + 3 + 4)	0
Official financing	0

Source: ONS, *Economic Trends*.

Table 24.1 shows the UK had a trade deficit in goods in 2011 Q4, offset partly by surpluses on trade in services. Combining trade in goods and services with net income from transfers, the **current account** of the balance of payments was £9 billion in deficit in the final quarter of 2011.

A current account surplus means that a country's foreign income exceeds its foreign spending. A current account deficit means that its foreign spending exceeds its foreign income. These surpluses and deficits are saving and dissaving, and lead to purchases or sales of foreign assets.

The *capital account* of the balance of payments records the international flows of transfer payments relating to capital items. This covers payments received from the EU regional development fund for investment in infrastructure projects, the transfer of capital into or out of the UK by migrants and the forgiveness of international debt by the UK government. In 2011 Q4 the UK had a net inflow of £1 billion on this capital account.

Table 24.1 shows a net financial inflow of £16 billion in 2011. The inflow of money to the UK as foreigners bought UK physical and financial assets exceeded the outflow of money from the UK as residents bought assets abroad. In part, this inflow reflected the fact that international investors thought that the UK was relatively safer than Europe at a time when the Eurozone was in crisis.

The balancing item, a statistical adjustment, would be zero if all previous items were correctly measured. It reflects a failure to record all transactions in the official statistics. Estimating implicit changes in the value of foreign investments – which the statistics treat as money brought home and then reinvested abroad again – is particularly tricky. The balancing item was quite large in 2011 Q4, showing that the statistics were otherwise mismeasuring to quite a large degree. Adding together the current account (1), the capital account (2), the **financial account** (3) and the adjustment (4), we obtain the UK balance of payments. It so happens that it just balanced in 2011 Q4.

The **financial account** of the balance of payments records international purchases and sales of financial assets.

The balance of payments shows the net inflow of money to the country when individuals, firms and the government make the transactions they wish to undertake under existing market conditions. It is in surplus (deficit) when there is a net inflow of money (outflow of money). It takes account of the transactions that individuals wish to make in importing and exporting and in buying and selling foreign assets, and the number of transactions that governments wish to make in the form of foreign aid (transfer payments to foreigners), military spending (maintaining military bases abroad), and so on.

The final entry in Table 24.1 is *official financing*. This is always of equal magnitude and opposite sign to the balance of payments in the line above, so that the sum of all six entries in Table 24.1 is *always* zero. Official financing measures the international transactions that the government must take to *accommodate* all the other transactions shown in the balance of payments accounts.

Floating exchange rates

If the exchange rate floats freely, there is *no* government intervention in the forex market. Forex reserves are constant. The exchange rate adjusts to equate the supply of pounds and the demand for pounds in the forex market.

The supply of pounds reflects imports to the UK and UK purchases of foreign assets. These are the outflows in the UK balance of payments

accounts. Conversely, the demand for pounds reflects UK exports and sales of UK assets to foreigners. These are the inflows in the UK balance of payments accounts. With a freely floating exchange rate, the quantities of pounds supplied and demanded are equal. Hence inflows equal outflows and the balance of payments is exactly zero. There is no intervention in the forex market and no official financing.

Since the balance of payments is the sum of the current account and the capital and financial accounts, under floating exchange rates a current account surplus must be exactly matched by a deficit on capital and financial accounts, or vice versa. This just says any unspent surplus on goods and services must be spent buying assets. A foreign deficit is financed by running down *net* foreign assets (lower assets or higher debt).

Fixed exchange rates

With a fixed exchange rate, the balance of payments need not be zero. When there is a deficit, total outflows exceed total inflows on the combined current and capital accounts. How is the deficit financed?

Since there is a deficit, the supply of pounds to the foreign exchange market, reflecting imports or purchases of foreign assets, exceeds the demand for pounds, reflecting exports or sales of assets to foreigners. The balance of payments deficit is exactly the same as the excess supply of pounds in the forex market.

To maintain the fixed exchange rate, the central bank offsets this excess supply of pounds by demanding an equivalent quantity of pounds. It runs down the foreign exchange reserves, selling dollars to buy pounds. In the balance of payments accounts this shows up as ‘official financing’.

When there is a balance of payments surplus, the government intervenes in the forex market to buy foreign exchange reserves. When there is a balance of payments deficit, reserves must be sold. Table 24.2 summarizes this discussion.

Table 24.2 Balance of payments and exchange rate regimes

Fixed exchange rate	Floating exchange rate
current account	current account
+ capital account	+ capital account

1 financial account	+ financial account
= balance of payments	= balance of payments
= [- official financing]	= 0
= rise in forex reserves	no official financing; no change in forex reserves

24.4 The real exchange rate

In 1981 the \$/£ exchange rate was \$2.03/£; by late 2012 it was only \$1.60/£. A fall in the international value of sterling makes UK goods cheaper in foreign currencies and foreign goods more expensive in pounds. Other things equal, the UK became more competitive as sterling fell.

The **real exchange rate** is the relative price of goods from different countries when measured in a common currency.

But other things were not unaltered. The UK had more inflation than the US, so its prices rose more during 1981–2012. UK competitiveness rose because of a lower nominal or actual exchange rate, but fell because the sterling price of UK goods rose by more than the dollar price of US goods. As usual, we must distinguish between nominal and real variables.

Thus if $E^{$/£}$ is the nominal exchange rate, measured by \$/£ the international value of sterling, and $p_{UK}^{\$}$ are $p_{US}^{\$}$ the domestic sterling price of UK goods and the dollar price of US goods,

$$\text{Real exchange rate} = \{E^{$/£} \times p_{UK}^{\$}\} / p_{US}^{\$} \quad (1)$$

Table 24.3 gives some examples. Pretend the only good is shirts. In row 1, a US shirt costs \$10 and a UK shirt £6. At a nominal exchange rate of \$2/£, the relative price of UK to US shirts, in a common currency, is 1.2, whether we compare the relative dollar price of shirts (\$12/\$10) or the relative price in pounds (£6/£5). Two things can make UK shirts more competitive in relation to US shirts.

Table 24.3 Calculating real exchange rates

Nominal exchange rate (\$/£)	UK shirt price (£)	UK shirt price (\$)	US shirt price (\$)	Real exchange rate
2.0	6	12	10	1.2
1.5	6	9	10	0.9
2.0	4.5	9	10	0.9
2.0	6	12	13.3	0.9

In row 2, a lower nominal exchange rate for sterling of \$1.50/£ reduces the relative price of UK to US shirts from 1.2 to 0.9. The UK's real exchange rate depreciated in equation (1) and the UK became more competitive since its shirts became cheaper when measured in a common currency.

In row 3, the nominal exchange rate is \$2/£, as in row 1, but now the sterling price of UK shirts has fallen from £6 to £4.50. At a nominal exchange rate of \$2/£, a UK shirt costs \$9. Since a US shirt costs \$10, the UK real exchange rate has again fallen to 0.9. Row 4 shows that a change in US prices can have the same result.

Equation (1) makes clear that the arithmetic of real exchange rates does not care whether the nominal exchange rate E falls, the sterling price of UK shirts falls or the dollar price of US shirts rises. Any one of these changes reduces sterling's real exchange rate and makes the UK more competitive (and the US less competitive). Conversely, a rise in the nominal exchange rate, a rise in UK sterling prices or a fall in US dollar prices increases sterling's real exchange rate and makes the UK less competitive (and the US more competitive).

Table 24.4 shows how this works out in practice. The first row shows that the nominal \$/£ exchange rate depreciated during 1981–2012. The second and third rows show what happened to the price level in each country. Setting the price index in each country equal to 100 in 1981, US prices rose from 100 to 252 by 2012, whereas UK prices had risen to 328.

The fourth row calculates an index of the real exchange rate, using the formula of equation (1). While the nominal exchange rate depreciated from 2.03 to 1.60 between 1981 and 2012, the real exchange rate index appreciated from 2.03 to 2.08. Despite the nominal exchange rate depreciation, the UK actually became less competitive in real terms because its prices rose by significantly more than those in the US.

Table 24.4 Nominal and real exchange rates

	1981	2012
\$/£	2.03	1.60
Prices (1981 = 1)		
UK (in £)	1.00	3.28
US (in \$)	1.00	2.52
Real \$/£ rate index (1981 = 2.03)	2.03	2.08
PPP exchange rate index (1981 = 2.03)	2.03	1.55

Source: IMF, *International Financial Statistics*.

Purchasing power parity (PPP)

What hypothetical path would the nominal exchange rate have had to follow to keep the real exchange rate at its initial level? The PPP exchange rate offers a quick check that lets us compare the present with what we know about the past.

The **purchasing power parity (PPP)** exchange rate path is the path of the nominal exchange rate that maintains a constant real exchange rate.

The final row of Table 24.4 shows what would have had to happen to the nominal exchange rate in order to maintain the real exchange rate at its 1981 level. If the 2012 nominal exchange rate had been \$1.53/£, this would exactly have offset the differential evolution of prices in the UK and US, restoring the real exchange rate to its original level.

Frequently asked questions

1. *Does it matter which is the base year in calculating a real exchange rate index?*

As with any index, there are many possible ways in which to construct it. Table 24.4 begins at 1981 and then goes forward. We could instead have chosen 2012 as the base year and worked backwards, asking how real exchange rates were different in 1981.

2. Does it matter which real exchange rate is used as the basis for computing the PPP path of the nominal exchange rate?

Same answer. The PPP path plots the path of a nominal exchange rate to accomplish a particular constant real exchange rate.

Choosing a different real exchange rate to be held constant would imply a different PPP path for the nominal exchange rate.

24.5

Determinants of the current account

Having defined the real exchange rate and discussed its relationship to competitiveness, we can now study what determines the current and capital accounts of the balance of payments. We begin with the current account.

Exports

Chapter 17 assumed that demand for exports was given. We now recognize that the demand for UK exports depends chiefly on two things. First, since UK exports are imports by the rest of the world, higher income abroad leads to higher UK exports. Second, a lower UK real exchange rate implies greater UK competitiveness and larger UK exports for any particular level of foreign income.

Exports respond quickly to changes in world income, but changes in competitiveness affect exports more slowly. Exporters may be unsure if the change in competitiveness is temporary or permanent. If they believe it to be temporary, they may change their profit margins but leave the price of their goods in foreign currency unaffected.

Even where this means losses in the short run, it may be cheaper in the long run than temporarily withdrawing from those markets and having to spend large sums on advertising and marketing to win back market share when competitiveness improves again. But if competitiveness fails to improve and the real exchange rate remains high, firms will gradually conclude that they should quit the exporting business.

Imports

Import demand is larger the higher is domestic income, as we recognized in Chapter 17 through the marginal propensity to import. But import

demand also rises when the real exchange rate depreciates making foreign goods cheaper relative to domestic goods when both are measured in the prices of the domestic currency. Again, in practice, imports respond more quickly to changes in domestic income than to changes in the real exchange rate. However, if sustained, an appreciation of the real exchange rate eventually raises imports.

Other items on the current account

Foreign aid and spending on military bases abroad are matters of government policy. The net flow of interest, dividend and profit income between countries arises because residents of one country hold assets in another. The size of this net flow of income depends on the pattern of international asset-holding and on the level of interest rates, profits and dividends at home and abroad.

24.6 The financial account

We have distinguished between transfer payments on the capital account, for example EU subsidies for road-building, and movements of financial capital to buy and sell assets on the financial account. The former are tiny and henceforth we ignore them completely, implicitly assuming that the capital account is in balance. However, financial flows on the financial account can be huge. These flows of financial capital are often called ‘capital flows’ even though they relate to the financial account.

Capital inflows and outflows reflect sales and purchases of foreign assets. These flows have become increasingly important. Computers and telecommunications make it as easy for a British resident to transact in the financial markets of New York or Frankfurt as in London. Moreover, controls on international capital flows have gradually been dismantled as a result of globalization and financial integration.

The world’s financial markets now have two crucial features. First, restrictions have been abolished for capital flows between the advanced countries. Funds can be freely moved from one country to another in search of the highest rate of return. Second, trillions of pounds are internationally footloose, capable of being switched between countries and currencies when assets in one currency seem to offer a higher rate of return than assets elsewhere.

Since the stock of international funds is now huge, capital flows could swamp the typical current account flows from imports and exports.

In international asset markets, capital gains arise not merely from changes in the domestic price of an asset but also from changes in exchange rates while temporarily holding a foreign asset – **speculation**. In Table 24.5, you can invest £100 for a year. UK interest rates are 10 per cent a year. US interest rates are zero. Keeping your funds in pounds, row 1 shows that you have £110 at the end of the year.

Speculation is the purchase of an asset for subsequent resale, in the belief that the total return – interest plus capital gain – exceeds the total return on other assets.

Table 24.5 Lending £100 for a year

£100 lent in:	Interest rate (%)		Exchange rate (\$/£)		Final wealth	
	UK	US	Initial	Final	\$	£
UK	10	-	-	-	-	110
US	-	0	2.0	1.8	200	110

CASE 24.1

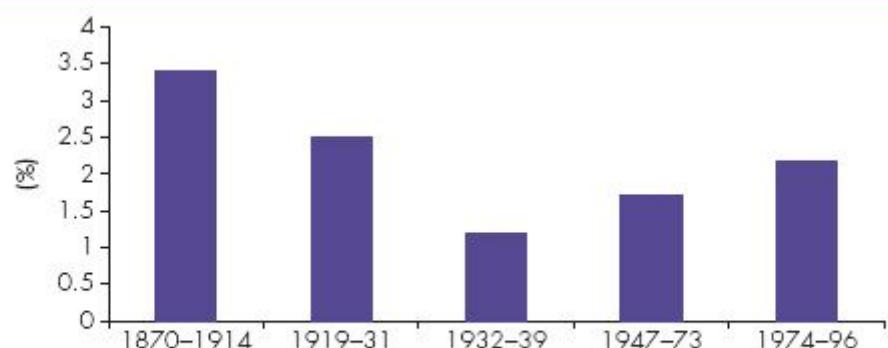
INTERNATIONAL FLOWS OF FINANCIAL CAPITAL

Flows on the financial account of the balance of payments may be short term, such as putting money in a foreign bank account, or long term, such as taking a permanent stake in a foreign company – **foreign direct investment (FDI)**.

Foreign direct investment (FDI) is the purchase of foreign firms or the establishment of foreign subsidiaries.

Has globalization made capital flows more important recently? The figure below shows the scale of average annual capital flows, relative to GDP, for 12 OECD economies in peacetime years during 1870–1996. Capital flows dried up in the 1930s, during the Great

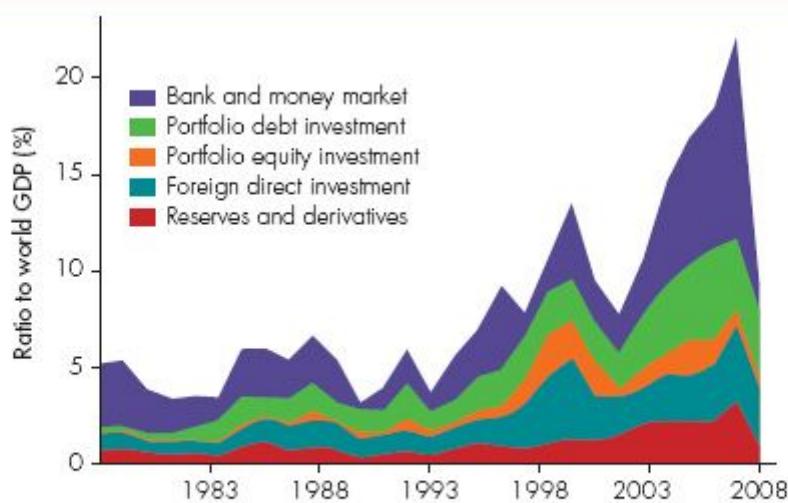
Depression, but today we tend to forget that the late nineteenth century was also a period of extensive foreign investment.



The scale of capital flows, 1870–1996 (% of GDP)

Source: M. Obstfeld, 'The global capital market: Benefactor or menace?', *Journal of Economic Perspectives* 12 (1998): 9–30.

Whereas this 100-year view shows no sign of an upward trend, a more recent focus shows a rise: since 1995, there was a four-fold rise in gross financial flows relative to world GDP, which itself was growing quickly. We now live in a world of highly mobile financial capital.



Gross international capital flows, 1983–2008

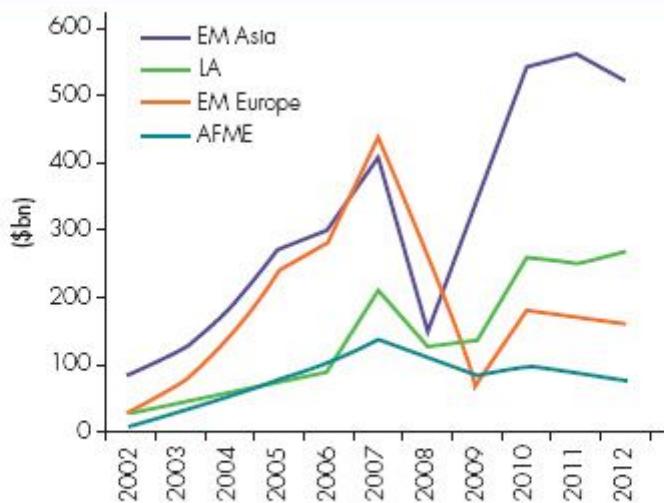
Source: C. Becker and C. Noone, *Volatility in international capital movements*, Reserve Bank of Australia.

You might have expected this to be one-way traffic: rich, advanced countries investing in emerging markets, such as China, India and

Brazil. This was not always the case. China has used its export surpluses to buy debt issued by Western governments, and also to lend to private sectors; 'sovereign wealth funds' from the oil-rich states of Abu Dhabi, Qatar, Kuwait and Bahrain have done the same. Without capital inflows on this scale, the pre-crash credit binge in Western countries would not have been as extensive.

The figure also confirms that international capital flows are increasingly volatile. This leads to two questions. First, when countries borrow from foreigners can they rely on this inflow being stable, or do they have to worry about possible outflows again? Second, if international capital is so mobile, could we, and should we, regulate it to reduce its mobility? These questions lie at the heart of international macroeconomics. We return to them in more detail in subsequent chapters.

The two previous figures need careful interpretation. They refer to gross flows – total inflows, total outflows, or the sum of the two. This is not the same as net inflows or outflows. Since official financing is usually small, if countries cannot run large current account deficits or surpluses, they cannot have large net capital flows either. In equilibrium, the size of the net flow must be of the same order of magnitude as the size of the current account. Since current account imbalances rarely exceed 10 per cent of GDP, we need to understand the market forces or policy responses that ensure that net capital flows are similar in size. The figure below shows the scale and volatility of net capital inflows to four emerging market regions – Emerging Europe, Emerging Asia, Latin America and Africa/Middle East. Capital inflows dried up a lot during the financial crisis but have subsequently resumed.



Net capital inflows

Source: Institute for International Finance, October 2012.

Finally, looking at the *size* of capital flows does not itself tell us about capital mobility, which relates to the *sensitivity* of capital flows to perceived profit opportunities. If exchange rates adjust to *prevent* massive capital flows, we will never see large flows in the data, whatever the degree of capital mobility.

Row 2 of Table 24.5 shows what happens if you convert £100 into dollars at an initial exchange rate of \$2/£, then lend this \$200 for a year at zero interest, to get \$200 by the year end. Suppose sterling depreciates 10 per cent during the year. At the year end, at the exchange rate of \$1.80/£, a fall of 10 per cent on the original rate of \$2/£, your \$200 converts back to £110. You get 10 per cent less interest than staying in the UK, but make a capital gain of 10 per cent by temporarily holding dollars, whose value relative to pounds rises 10 per cent in the year.

In this example you end up with £110 whether you lend in dollars or in pounds for the year. If the pound depreciates more than 10 per cent, the capital gain on holding dollars outweighs the loss of interest, and the total return on lending in dollars is higher than in pounds. Conversely, if the pound depreciates against the dollar by less than the interest rate differential, you earn a higher total return by keeping your money in pounds.

Equation (2), which is called the **interest parity** condition, summarizes this important result. The total return on temporarily lending in a foreign

currency is the interest rate paid on assets in that currency plus any capital gain (or minus any capital loss) arising from depreciation (appreciation) of the domestic currency during the period.

Interest parity means that expected exchange rate changes offset the interest differential between domestic and foreign currency assets.

$$\text{Return on domestic asset} = \text{return on foreign asset} \quad (2)$$

$$= \text{foreign interest rate} + \% \text{ depreciation of exchange rate while funds abroad}$$

With near **perfect capital mobility**, there is a vast capital outflow if the total return on foreign lending exceeds the total return (the domestic interest rate) on domestic lending. There is a huge capital inflow if the return on domestic lending exceeds the return on lending abroad.

Perfect capital mobility means that a vast quantity of funds flows from one currency to another if the expected return on assets differs across currencies.

Net flows on the financial account of the balance of payments are small only when the total return on foreign lending is similar to the return on lending in the domestic currency. With no barriers to capital mobility, expected total returns are the same in assets of different currencies. Expectations about the future determine the capital gains or losses that people expect to make through changes in the exchange rate.

MATHS 24.1

INTEREST PARITY CONDITIONS

Let r denote the domestic interest rate, r^* the foreign interest rate, s the nominal exchange rate (international value of the domestic currency) and ds/dt the instantaneous rate of change of the exchange rate. $(1/s)ds/dt$ is then the instantaneous percentage capital gain that a foreign investor makes by holding the domestic currency for an instant before repatriating the money, and, under perfect certainty, the interest parity condition implies

$$r^* = (1/s)ds/dt + r \quad (1)$$

The real exchange rate v is given by

$$v = sp/p^* \quad (2)$$

where p is the domestic price level and p^* the foreign price level. The instantaneous change in the real exchange rate obeys

$$(1/v)dv/dt = (1/s)ds/dt + (1/p)dp/dt - (1/p^*)dp^*/dt = (1/s)ds/dt + \pi - \pi^* \quad (3)$$

Real exchange rate appreciation reflects nominal exchange rate appreciation, domestic (p) inflation or foreign deflation ($-\pi^*$).

Combining equations (1) and (3):

$$\begin{aligned} r^* - \pi^* &= r - \pi + (1/s)ds/dt + \pi - \pi^* \\ \Rightarrow [r^* - \pi^*] &= [r - \pi] + [(1/v)dv/dt] \end{aligned} \quad (4)$$

Thus, the interest parity condition expressed in nominal terms in equation (1) – nominal interest differentials must be offset by appropriate capital gains or losses in nominal exchange rates to preserve the equality of return in different currencies under perfect international capital mobility – implies a similar statement in terms of real interest rate differentials being offset by capital gains or losses on the real exchange rate.

Although derived for an instantaneous decision, we can always view a longer horizon as a series of instant decisions. Hence, interest parity conditions also hold over longer horizons, provided the duration of the interest rates matches the period over which exchange rate changes are assessed.

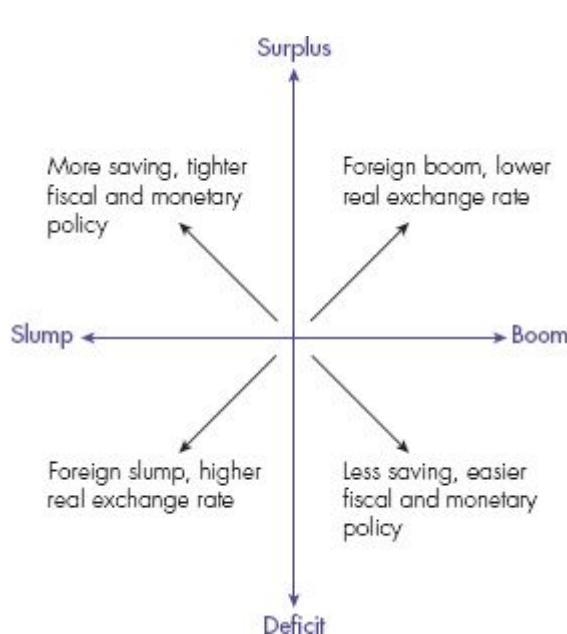
Finally, once uncertainty exists, we have to replace actual exchange rate changes by those expected at the outset of the period. An investor contemplating lending abroad for a year can always obtain a one-year foreign bond with a known interest rate today, but will have to take a view on the likely change in the exchange rate over the year.

24.7 Internal and external balance

Next, we discuss the relationship between the state of the economy – boom or recession – and the current account on the balance of payments.

Figure 24.3 shows the different combinations of boom and recession and current account surpluses and deficits. Think about demand and supply for domestic output. Equation (3) reminds us of the basic equation for goods market equilibrium:

$$Y = C + I + G + (X - Z) \quad (3)$$



With internal and external balance there is neither a boom nor a slump, and the current account just balances. Each quadrant of the diagram identifies shocks that cause departures from internal and external balances. For example, tight fiscal and monetary policy reduces aggregate demand, creating a domestic slump but a current account surplus since import demand is reduced. However, by increasing export demand, a foreign boom leads to both a domestic boom and a current account surplus. Other possible shocks and their consequences are shown.

Figure 24.3 Internal and external balance

Domestic output Y equals aggregate demand that arises from spending on consumption, investment, government purchases and net exports. If aggregate demand for domestic output equals potential output, firms

produce the full-employment output level and in the labour market demand as much employment as workers wish to supply.

With sluggish wage and price adjustment, lower aggregate demand causes a recession. Only when aggregate demand returns to potential output is **internal balance** restored.

A country is in **internal balance** when aggregate demand equals potential output.

For a floating exchange rate, the total balance of payments is always zero. Saying that the current account is in internal balance then also implies financial account balance.

In Figure 24.3 the point of internal *and* **external balance** is the intersection of the two axes, with neither boom nor slump, and with the current account in neither surplus nor deficit.

A country in **external balance** has a zero current account balance.

Internal balance implies aggregate demand equals potential output, and there is full employment in the labour market. External balance means current account balance. The country is neither under-spending nor over-spending its foreign income. Nor is it augmenting or depleting its foreign assets. Foreigners are not acquiring domestic assets without limit, nor are domestic residents acquiring ever-larger holdings of foreign assets.

Figure 24.3 shows how shocks move the economy away from internal and external balance. For example, the top left-hand quadrant shows a combination of domestic slump and current account surplus. This can be caused by a rise in desired saving (a downward shift in the consumption function) or by the adoption of tight fiscal and monetary policy. These reduce aggregate demand and thereby cause both a domestic slump and a reduction in imports.

Similarly, a higher real exchange rate (lower competitiveness) reduces export demand and raises import demand. The fall in net exports induces both a current account deficit and lower aggregate demand, leading to a domestic slump as shown in the bottom left-hand quadrant. The figure shows other shocks that move the economy into other quadrants, causing departures from both internal and external balance.

A key lesson of Figure 24.3 is that most shocks in an open economy move the economy away from *both* internal *and* external balance. In studying a closed economy, we examined whether the economy could return to internal balance on its own. When adjustment is sluggish, monetary and fiscal policy can speed up adjustment. In a slump, expansionary monetary and fiscal policy hasten the return to full employment.

24.8 The long-run equilibrium real exchange rate

In **long-run equilibrium**, both internal and external balance must hold. Domestic output Y is at potential output Y^* and the current account is in balance. For countries with large foreign debts or foreign assets, and thus large flows of interest income, the current account can deviate a lot from the trade balance. However, for most countries, the trade balance and the current account balance are similar.

Simultaneous internal and external balance is the **long-run equilibrium** of the economy.

Initially, we focus on this latter case. External balance then requires that net exports $X - Z$ must be zero. Long-run equilibrium then requires

$$Y^* = Y = [C + I + G] + [X - Z] \quad (4)$$

In external balance, net exports $(X - Z) = 0$. Internal balance then requires that domestic demand $(C + I + G)$, the domestic absorption of resources, equals potential output Y^* .

Net exports depend on real income at home, real income abroad and the real exchange rate that determines competitiveness. In long-run equilibrium, both domestic and foreign income are fixed at their respective levels of potential output. Given these income levels, net exports depend only on the real exchange rate.

ACTIVITY 24.1

CHANGES IN EQUILIBRIUM REAL EXCHANGE RATES

Paul Samuelson, one of the fathers of modern economics, won the Nobel Prize for his work on many aspects of economics, including international trade. Empirical research confirms a relation first noticed by Béla Balassa and Paul Samuelson: countries with higher per capita real incomes have a higher real exchange rate.

Typically, there is more technical progress in industries making goods for trade (computers, cars, telecommunications) than in industries making services for the home economy (haircuts, laundry, crèches). Similarly, productivity-enhancing capital accumulation occurs mainly in the traded goods sector. The main difference between a rich country and a poor country is not that hairdressers or childminders are more productive in rich countries, but that industries making exports and competing with imports are more productive.

Countries with high per capita incomes therefore have high real exchange rates because their traded goods sector is more productive. Without real exchange rate appreciation such countries would be too competitive. Why does the Balassa–Samuelson effect matter? Here are some examples.

At what exchange rate should Eurozone members be admitted?

The Eurozone is a monetary union whose members have permanently fixed exchange rates against one another but a floating exchange rate in relation to the rest of the world. Suppose, just suppose, the Eurozone decided to allow Turkey to join their fixed exchange rate club. A country outside the Eurozone will typically have a floating exchange rate which will move around a bit. How do negotiators decide on a sensible exchange rate to permanently fix to the currency of a new entrant?

They might look at past data, hoping to find a period of internal and external balance in that country. This would be a starting point for calculating a sensible real exchange rate that would provide just the right amount of competitiveness.

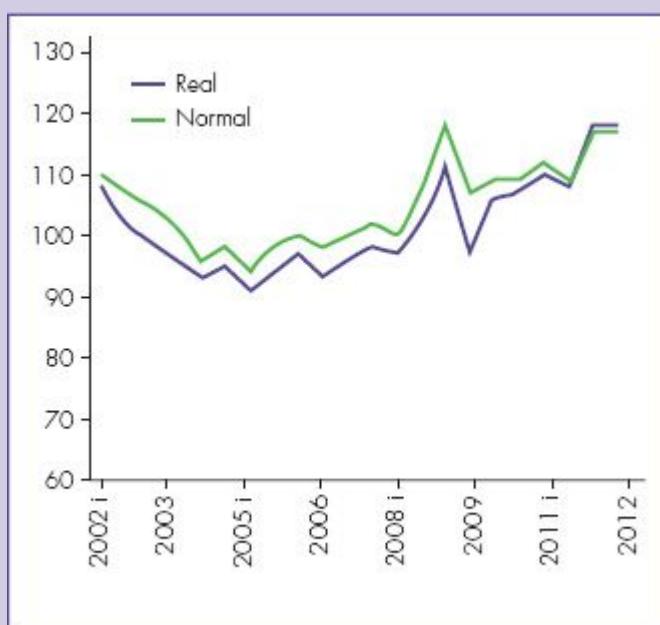
So, from the base date, they would calculate how much relative prices of Turkey and the Eurozone had changed up to the date of

Turkish entry, and adjust the original nominal exchange rate by this amount to restore the real exchange rate to the level at which it had last appeared correct. This would determine the nominal exchange rate at which Turkey was admitted to the common currency.

Without knowing about the Balassa–Samuelson effect, negotiators might make a mistake in assuming that the past was the perfect guide to the future. More sophisticated negotiators might take the above as a starting point but then ask whether Turkey was still an emerging economy, in relation to the more mature Eurozone economies, and therefore make an estimate of the further real appreciation of the Turkish currency that might be compatible in the longer run with achieving a sustainable real exchange rate. This would lead to a different estimate of a suitable initial exchange rate for Turkish entry.

Over-competitive China

Perhaps the most obvious example in the current global economy is the Chinese exchange rate. For years, China chose to fix its exchange rate to the US dollar, and to maintain this peg at a level that kept the Chinese economy super-competitive: China had massive current account surpluses (and hence also an outflow of capital as it invested these surpluses abroad again). On 23 February 2010 the *Financial Times* reported an estimate that the Chinese exchange rate was undervalued by 41 per cent relative to the level consistent with internal and external balance.



Source: Financial Times, 15 April 2013

The figure shows China's real exchange rate since the first half of 2002. During 2005–12 there was a nominal and real appreciation of over 20 per cent. At first sight, this considerably reduces Chinese competitiveness but, wearing Balassa–Samuelson spectacles, productivity in China also grew rapidly during this period. Output grew by 10 per cent a year, and the one-child policy meant that population growth was negligible – output per person thus grew rapidly, preserving competitiveness despite the appreciating real exchange rate.

Policy makers, from Washington to Paris, complain that China's exchange rate policy is bankrupting their economies, leading to an export of jobs from the West to Asia, and leaving Western economies exposed to the inflows of financial capital from China that might, at some future date, decide to become outflows.

A faster appreciation of China's nominal exchange rate peg would help but it is not the only adjustment mechanism. Chinese domestic prices and wages may rise sufficiently more quickly than those in the West to achieve further real appreciation – though the figure above suggests this has not been important so far.

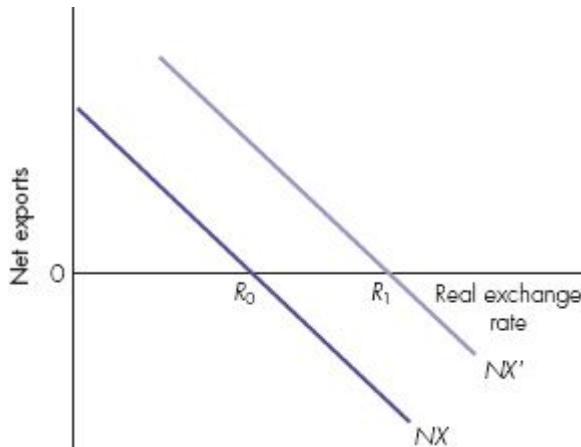
Questions

- a. Except for footballers, investment bankers and university professors, labour is largely a commodity that is not traded across national frontiers. Does this mean that countries with a high real exchange rate will also be those with a high real wage rate?
- b. Rank the following countries in terms of the current level of their real exchange rate, with the highest first: China, Greece, Italy, Switzerland. Which country should have the least scope for real appreciation in the long run?
- c. Suppose the Internet allows extensive international trade in services (for example, legal, accounting, software and entertainment services). Is the Balassa–Samuelson effect then likely to break down? Why, or why not?

To check your answers to these questions, go to page 681.

Figure 24.4 shows that there is a unique real exchange rate that makes net exports equal to zero. Given domestic and foreign levels of potential output, a lower real exchange rate raises export demand and reduces import demand. The net export schedule NX slopes down. Only at the real exchange rate R_0 are net exports zero. At a higher real exchange rate, competitiveness is too low and net exports are negative. At a lower real exchange rate, competitiveness is too high and net exports are positive.

Beginning from R_0 , suppose the country gets a favourable and lasting supply shock that raises potential output Y^* . For example, the country discovers a natural resource, such as oil or gold, or develops a new high-tech industry, such as computers. Since the marginal propensity to consume is less than unity, if output and income rise by 100, aggregate demand rises by less than 100. The remaining output is exported and net exports rise.



Given domestic and foreign incomes, a higher real exchange rate reduces competitiveness and net exports. Only at R_0 is there trade balance. A resource discovery, such as North Sea oil, shifts NX to NX' causing an appreciation of the real exchange rate to R_1 to maintain trade balance in the long run.

Figure 24.4 The long-run equilibrium real exchange rate

In Figure 24.4 the favourable supply shock shifts the net export schedule to NX' and the long-run equilibrium real exchange rate appreciates from R_0 to R_1 . If finding North Sea oil adds to UK net exports, only a fall in the country's manufacturing exports will prevent a permanent trade

surplus. A real exchange rate appreciation – a fall in UK competitiveness – is the market mechanism that restores external balance.³

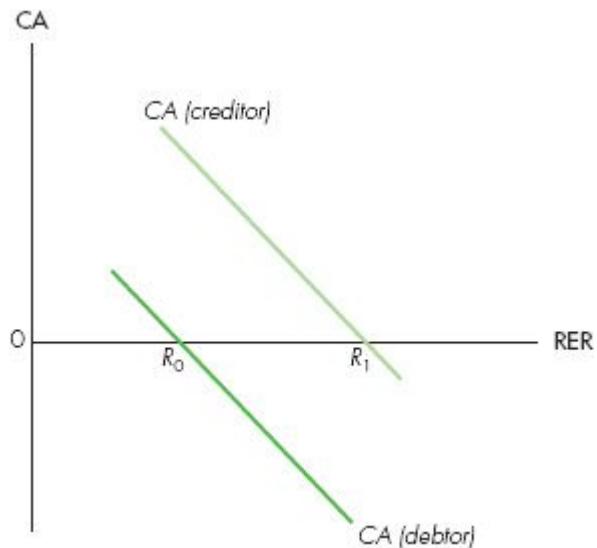
Large supply shocks, such as a big resource discovery, are the exception not the norm. If no shocks occur, the real exchange rate is constant in long-run equilibrium. This has two implications. First, if domestic and foreign prices grow at different rates, the nominal exchange rate has to adjust steadily to keep the real exchange rate constant. The nominal exchange rate then follows the purchasing power parity path discussed in Section 24.4.

Second, if the nominal exchange rate is fixed as an act of policy, it is possible to maintain a constant real exchange rate in the long run only if domestic and foreign prices change at the same rate. Otherwise the real exchange rate is changing in the long run, and net exports will not remain zero, as external balance requires.

Foreign debt and foreign assets

Finally, we recognize that some countries have important flows of international income or payments as a result of owning large foreign assets or having large foreign debts. The current account is net exports ($X - Z$) plus rA ; that is, the stock of net foreign assets multiplied by the interest rate r . For creditor countries A is positive; for debtor countries A is negative.

Figure 24.5 shows how inherited foreign assets or debts affect the long-run equilibrium real exchange rate. The current account CA is net exports NX , as in Figure 24.4, plus net interest on foreign assets. In order to achieve current account balance, a debtor country needs a low real exchange rate R_0 to be competitive and a sufficient trade surplus to pay interest on its foreign debts. A creditor country has a high real exchange rate R_1 to reduce competitiveness and run a trade deficit, financed by interest earned on foreign assets.



The current account CA is net exports NX plus net interest on foreign assets. For current account balance, a debtor country needs a low real exchange rate R_0 to be competitive and have a sufficient trade surplus to pay interest on its foreign debts. A creditor country has a high real exchange rate R_1 to reduce competitiveness and run a trade deficit, financed by interest earned on foreign assets.

Figure 24.5 Foreign assets and the real exchange rate

This figure is helpful in thinking about implications of foreign indebtedness in the aftermath of the financial crisis. Because of its large financial sector, the UK had above-average exposure to the financial crash. The UK government's debt increased sharply as it bailed out failing financial institutions. Figure 24.5 implies that, to the extent that interest payments are made to foreign investors, the real sterling exchange rate should permanently depreciate as a result of the crash. To service permanent interest payments to foreigners, the UK has to run a larger trade surplus than before, which needs greater competitiveness achieved by a lower exchange rate.

What about countries, such as Greece, in the Eurozone? With much higher debts than previously recognized, Greece requires a real depreciation, for the same reason as the UK. This could be achieved by a nominal depreciation of the euro. But the Eurozone also includes some healthier economies that do not need a depreciation. The textbook solution therefore includes a fall of domestic prices in Greece in order to improve its competitiveness by depreciating its real exchange rate.

This requires a temporary period of lower aggregate demand in Greece than elsewhere. Fiscal contraction not merely deals with its budget deficit, but also reduces aggregate demand, putting downward pressure on prices and wages in Greece. With sluggish wage adjustment in the Greek labour market, this is a recipe for high unemployment while adjustment is taking place. Whether the Greek government is strong enough to administer the medicine is something that bond markets are watching with concern.

This completes our analysis of the long-run equilibrium exchange rate, compatible with both internal and external balance. In the long run, it is the current account of the balance of payments that affects the exchange rate. The financial account gets into the story only to the extent that the cumulation of *past* capital flows is what determines the current stock of net foreign assets.

In the short run, the story is very different. Countries can run current account surpluses and deficits. Short-run changes in the exchange rate then have much more to do with the financial account. The role of capital flows is one theme of the next chapter. The other themes are how the economy adjusts to temporary shocks and returns to internal and external balance, whether macroeconomic policy can ease this adjustment and how choice of exchange rate regime affects these issues.

CASE 24.2

CURRENCY WARS

When times are good and life is easy, it is not too hard to behave well. When the going gets tough, good behaviour is harder to sustain.

During the depression of the 1930s, countries engaged in 'competitive exchange rate depreciations'. Since every exchange rate is a relative price, one country's depreciation and additional competitiveness is another country's appreciation and reduced competitiveness. The world as a whole cannot enhance competitiveness, which is necessarily competitiveness against trading partners.

When a currency is agreed to be overvalued, its depreciation may be widely welcomed since it restores more sensible and sustainable

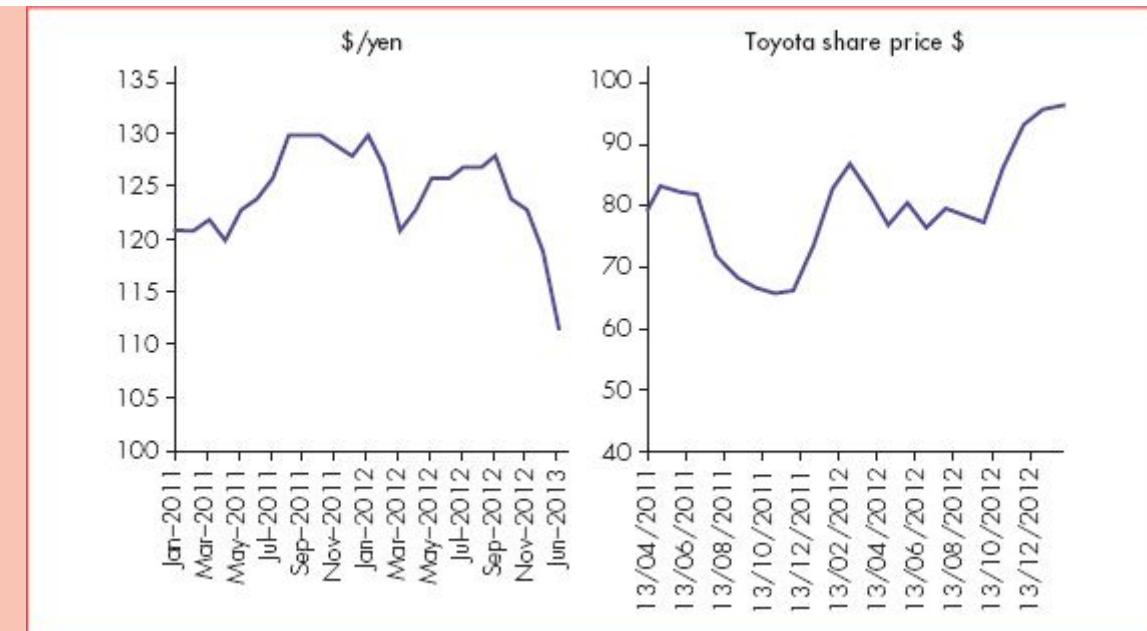
levels of the exchange rate. However, when worldwide aggregate demand is insufficient, for whatever reason, attempts by one country to gain aggregate demand *at the expense of others* by depreciating its exchange rate will be unwelcomed by its trading partners. This is sometimes called an attempt to export unemployment.

During the 1930s, as the depression intensified, the countries that depreciated first gained an advantage. Even when their trade partners retaliated with a depreciation of their own, this rarely did more than erode the advantage that had initially been gained. Moreover, there was nothing to stop the original countries engineering another exchange rate depreciation. The eventual result was considerable volatility and uncertainty that hampered trade flows, to everyone's detriment. One of the reasons that the world adopted a fixed exchange rate system in 1945 was a desire to avoid another set of unhelpful currency wars.

By the late 1990s, most countries had adopted independent central banks that set interest rates in pursuit of some kind of inflation target. Although the motivation for this was inflation stability, it had an important, though largely unintended, consequence.

Since different countries adopted very similar inflation targets, they ended up with very similar monetary policies. With both inflation and interest rates relatively similar and stable across countries, there were few pressures for substantial exchange rate changes. Those currency tensions that did emerge – we have already discussed concerns about over-competitive China – were as much the result of long-term changes in Chinese productivity as the result of any short-term currency manipulation.

The longer that stagnation and austerity continue in the wake of the financial crash, the more sensitized are policy makers to attempts by any country to engage in competitive depreciation. In 2012, after 20 years of stagnation and deflation, Japan finally elected a government more resolutely committed to printing money, raising the inflation rate and, by implication, allowing the Japanese yen to depreciate. The figure below shows how markets steadily increased their belief that a new policy was indeed under way.



After a period of exchange rate stability, the yen fell substantially against the US dollar and other currencies. Despite the fact that the inflation target had been increased by only 1 percentage point a year, the international value of the yen fell by more than 15 per cent.

For Japanese exporters such as Sony and Toyota, this was extremely welcome news. The right-hand chart shows how Toyota's share price (in US dollars) has risen as Japanese competitiveness has improved. What has been good news for Toyota has been less welcomed not only by Ford, GM and Volkswagen, but also by nearer neighbour Hyundai.

To date, there has been little opposition to Japanese policies to depreciate the yen. Japan is such a large country that the world economy will be more prosperous if Japan finally resumes economic growth (and hence imports from other countries). However, if the depreciation proceeds too far, and Japan becomes too competitive, greater opposition is bound to emerge.

Summary

- The **exchange rate** is the number of units of foreign currency that exchange for a unit of the domestic currency. A fall (rise) in the

exchange rate is called **depreciation (appreciation)**.

- The **demand for domestic currency in the forex market** arises from exports and purchases of domestic assets by foreigners; the **supply of domestic currency** to the market arises from imports and purchases of foreign assets. **Floating exchange rates** equate supply and demand for currency in the absence of government intervention in the forex market.
- Under **fixed exchange rates**, the government meets an excess supply of pounds by running down foreign currency reserves in order to prompt demand for pounds. An excess demand for pounds, at the fixed exchange rate, raises the foreign exchange reserves as pounds are supplied to the market.
- In the **balance of payments accounts**, monetary inflows are credits and monetary outflows are debits. The **current account** shows the trade balance plus current transfer payments, which largely reflect income earned from assets owned in other currencies, payment of international subsidies and social security payments. The capital account records the transfers of capital by migrants, debt forgiveness and net grant receipts for infrastructure projects from overseas institutions. Typically, this is small and for convenience we often ignore it completely. The **financial account** shows net purchases and sales of foreign assets. The balance of payments is the sum of the current, capital and financial account balances.
- Under floating exchange rates, a current surplus must be offset by a financial account deficit or vice versa. Under fixed exchange rates, a balance of payments surplus or deficit must be matched by an offsetting quantity of official financing. **Official financing** is government intervention in the forex market.
- The **real exchange rate** adjusts the nominal exchange rate for prices at home and abroad, and is the relative price of domestic to foreign goods

when measured in a common currency. A rise in the real exchange rate reduces the **competitiveness** of the domestic economy.

- The **purchasing power parity** is the path of the nominal exchange rate that would keep the real exchange rate at its initial level.
- An increase in domestic (foreign) income increases the demand for imports (exports). An increase in the real exchange rate reduces the demand for exports, increases the demand for imports and reduces the demand for net exports.
- Holders of international funds compare the domestic interest rate with the total return from temporary lending abroad. This return is the foreign interest rate plus the depreciation of the international value of the domestic currency during the loan. Perfect international capital mobility means that an enormous quantity of funds shifts between currencies when the perceived rate of return differs across currencies.
- The **interest parity** condition says that, when capital mobility is perfect, interest rate differentials across countries should be offset by expected exchange rate changes, so that the total expected return is equated across currencies.
- **Internal balance** means output is at potential output. **External balance** means the current account equals zero. Long-run equilibrium needs both.
- Given domestic and foreign levels of potential output, there is a unique real exchange rate that achieves trade balance. An increase in domestic potential output, for example from a resource discovery, causes a real exchange rate appreciation to maintain trade balance in the long run.
- Interest flows from foreign assets and debts make the current account differ from the trade balance. The higher are net foreign assets, the higher is the inflow of interest income and the higher is the real exchange rate needed to maintain external balance.

Review questions



EASY

- 1 If \$1 is traded for 1 euro and \$1.40 is traded for £1, what is the exchange rate between the euro and the pound sterling? Can the dollar appreciate against the euro but not against the pound?
- 2 Suppose the initial exchange rate is \$4/£. After 10 years, the US price level has risen from 100 to 300 and the UK price level has risen from 100 to 200. What nominal exchange rate would preserve purchasing power parity?
- 3 Which of the following statements is correct? (a) An exchange rate appreciation causes a loss of competitiveness. (b) If a country gained competitiveness for other reasons, such as a technological improvement, the consequence would be an appreciation of its equilibrium real exchange rate. (c) In the short run, exchange rates are driven more by the views of speculators than the need to balance imports and exports. (d) All of the above. (e) None of the above.
- 4 **Common fallacies** Why are these statements wrong? (a) Countries with lower inflation gain competitiveness. (b) Current and financial accounts are equally important in determining the level of floating exchange rates in the short run. (c) UK interest rates are high. This means the pound will appreciate for the next few months.
- 5 For decades, Japan has had a trade surplus. Must countries eventually get back to external balance? Is there more pressure on deficit countries than surplus countries to restore external balance?

MEDIUM

- 6 A country has a current account surplus of £6 billion but a financial account deficit of £4 billion. (a) Is its balance of payments in deficit or surplus? (b) Are its foreign exchange reserves rising or falling? (c) If the country has a fixed exchange rate, is the central bank buying or selling domestic currency? Explain.
- 7 Newsreaders say that ‘the pound had a good day’ if the sterling exchange rate rises. When is an appreciation (a) desirable and (b) undesirable?
- 8 The table below shows the price of a children’s toy in the UK and its price in the US. Use the data in the table to calculate the real exchange rate.

Nominal exchange rate	UK price of toy (£)	US price of toy (\$)
3.5	8	10
4	8	7
3.5	5.5	7
3.5	8	10

- 9 A country discovers oil and its real exchange rate appreciates. Manufacturers go bust because their exports are no longer competitive. Could the country be worse off as a result of finding this valuable resource?
- 10 Suppose Bob Geldof and Bono succeed in getting all the debts of poor countries written off. What happens to the real exchange rate of (a) the poor countries and (b) the rich countries? (c) What happens to the manufacturing exports of rich countries? (d) If there were single monopoly producers of manufactured goods in rich countries, how would they have been lobbying their governments?
- 11 Does Manchester have a balance of payments with everyone else? By what mechanism is long-run equilibrium achieved?

HARD

- 12 The following table shows country A's bilateral exchange rate against country B and country C. If the countries are equally important trade partners of country A, what is happening to country A's effective exchange rate? If it trades twice as much with country B as country C, what is the evolution of country A's effective exchange rate?

Year	2014	2015	2016
Country A's exchange rate index against:			
Country B	100	200	400
Country C	100	50	25

- 13 **Essay question** ‘Capitalist firms have no problem prospering despite the volatility of stock markets. Nobody has ever suggested government policies to fix stock market prices. Exchange rates are just another asset price and it is just as silly to fix exchange rates. Let them float.’ Why do governments ever want to fix exchange rates?

|4 Suppose Greece has to borrow extensively from foreign countries, thereby acquiring substantial foreign debt. Show graphically what has to happen to its equilibrium real exchange rate. Why is this change required? If Greece remains within the Eurozone, how can such a change be accomplished?

- 1 We are thus using the ‘international value of sterling’ as the measure of ‘the’ UK exchange rate.
- 2 The supply and demand for cars refers to physical quantities supplied or demanded at each price. However, the supply and demand schedules for pounds sterling refer to values of pounds supplied and demanded at each exchange rate. That is why the analysis can be more confusing than the analysis of the market for physical commodities. ‘Number of pounds’ on the horizontal axis is really a value not a quantity.
- 3 The fact that a resource discovery hurts other sectors, such as manufacturing, is sometimes called Dutch Disease. Holland’s real exchange rate appreciated significantly after its discovery of offshore gas fields in the North Sea. Sterling also appreciated after the UK subsequently found North Sea oil.

CHAPTER 25

Open economy macroeconomics

Learning outcomes

By the end of this chapter, you should be able to:

- 1 understand price and output adjustment under fixed exchange rates
 - 2 describe monetary and fiscal policy under fixed exchange rates
 - 3 explain the effects of devaluation
 - 4 understand what determines floating exchange rates
 - 5 describe monetary and fiscal policy under floating exchange rates
- Having introduced fixed and floating exchange rate regimes, we now study how the exchange rate regime affects the way in which an economy operates.

Having introduced fixed and floating exchange rate regimes, we now study how the exchange rate regime affects the way in which an economy operates.

Openness is often measured by the size of exports (or imports) relative to GDP. However, links through financial markets often have more impact. Large outflows of financial capital can provoke acute crises. Such crises may induce austerity measures to reassure foreign investors, devaluation of a pegged exchange rate or adoption of a completely new exchange rate regime.

UK discussions about future exchange rate policy still recall the day in 1992 that the UK was forced off a pegged exchange rate in the Exchange Rate Mechanism. Similarly, since the financial crash, the options for Greece, pegged to its Eurozone partners, have been very different from those of the UK, whose exchange rate remains freely floating. Even in the absence of crises, the choice of exchange rate regime affects the transmission mechanism of both monetary and fiscal policy. In this chapter, we study how our analysis for a closed economy must be amended for an [open economy](#).

Initially, we examine fixed exchange rate regimes. Then we discuss the determination of floating exchange rates and the consequences for macroeconomic policy.

25.1 Fixed exchange rates

Capital controls are regulations preventing private sector financial capital flows between different currencies.

The balance of payments and the money supply

Open economy macroeconomics examines how the economy is affected by links with other countries through trade, the exchange rate and capital flows.

To understand the role of capital mobility, suppose initially that there are no private sector capital flows. Most economies had **capital controls** during the period of global fixed exchange rates from 1945 to 1973. Gradual integration of global financial markets made these controls less effective. Once closer financial market integration became an explicit policy objective, capital controls were completely scrapped.

With a fixed exchange rate but no private capital flows, suppose the economy has a current account deficit and hence a balance of payments deficit. To finance the deficit, but preserve the exchange rate, the forex reserves must fall. The central bank sells foreign exchange and buys domestic currency, demanding the domestic currency that nobody else wants. In consequence, domestic money in circulation falls as pounds disappear back into the central bank. The balance of payments deficit reduces the domestic money supply. A balance of payments surplus would increase the money supply.

Under fixed exchange rates, the money supply is not determined exclusively by the original decision about how much domestic money to create. It also depends on the balance of payments surplus or deficit. When there is a payments surplus (deficit), cash flows into (out of) the country, directly changing narrow money, which in turn affects bank deposits and broad money.

Unsterilized intervention uses forex reserves to offset balance of payments imbalances. The exchange of foreign reserves for domestic cash alters cash in circulation and the domestic money supply.

Sterilization – an open market operation between domestic money and domestic bonds – offsets the change in domestic money supply that a balance of payments surplus or deficit would otherwise induce.

Allowing the balance of payments to change the money supply at a fixed exchange rate is called **unsterilized intervention** in the forex market. With a different money supply, interest rates have to change to adjust money demand by the same amount.

A country begins with a balance of payments deficit. Do induced changes in the money supply help the economy adjust to a long-run equilibrium in which external payments are in balance? This depends on the exact nature of the monetary policy in force.

First, suppose the central bank is pursuing a flexible inflation target. Its principal concerns are what is happening to inflation and to output. An external deficit may eventually impact on what is of direct concern to the central bank, but it is not an immediate priority. To achieve the interest rate desired, the central bank is passively adjusting the money supply all the time. If an external deficit induces a monetary outflow that would have undesired interest rate implications, the central bank automatically *undoes* this effect again – **sterilizing** the effect on the domestic money supply. The external deficit then makes no contribution to any adjustment mechanism.

Second, suppose the central bank is pursuing a target for the nominal money supply. Any balance of payments induced effect on the domestic money supply is immediately sterilized, as a matter of policy, to achieve the money supply target. Again, there is no automatic mechanism of adjustment through induced changes in the money supply.

Such a mechanism could arise if monetary policy commits *not* to sterilize the consequences of balance of payments imbalances. External deficits would reduce the money supply, force a rise in interest rates, reduce aggregate demand and choke off demand for imports. Such policies have, from time to time, been advocated by monetarists who see the evolution of monetary aggregates as critical to interpreting what is happening to the economy. However, this takes us a long way away from any monetary policies currently in operation in the main countries of the world.

Changes in money, and thus aggregate demand and income, may sometimes act as an adjustment mechanism to external imbalances. Later in the chapter, we examine how changes in inflation can affect competitiveness, and thereby the current account balance.

The role of capital mobility

So far, we have identified external imbalances only with trade surpluses or deficits. Now we introduce highly mobile private capital. If international investors have more funds at their disposal than central banks, central banks no longer defend exchange rates by buying and selling foreign exchange reserves.

Instead, central banks set domestic interest rates to provide the correct incentive for speculators.

A change in interest rates manipulates capital flows and the financial account of balance of payments. Since these flows can be huge, in the short run this dwarfs the current account of the balance of payments. Fixing the exchange rate is now a commitment to set the correct interest rate to eliminate one-way capital flows. This interest rate, coupled with the level of domestic income, determines money demand. This must equal real money supply. Given inherited prices, this determines the nominal money supply.

Thus, in the short run, only one level of the nominal money supply will do this job. Suppose the central bank tries further domestic open market operations between money and bonds. If it boosts the money supply, interest rates fall, capital flows out until the money supply falls back again, and interest rates return to the only level compatible with the pegged exchange rate.

When capital mobility is high, adjustment back to long-run equilibrium no longer occurs through induced changes in the money supply and interest rates.

The important conclusion is that pegging the exchange rate when capital mobility is very high means subordinating domestic interest rates to the single task of defending the exchange rate. There is no scope for any other choice of monetary policy. In the language of earlier chapters, the *LM* schedule is horizontal at the given interest rate necessary to maintain the pegged exchange rate. Similarly, there is no *ii* schedule or flexible inflation targeting. In such a world, there is little scope for an independent monetary policy committee. Choosing a fixed exchange rate means abandoning any other monetary policy. Interest rates are dedicated entirely to the exchange rate defence.

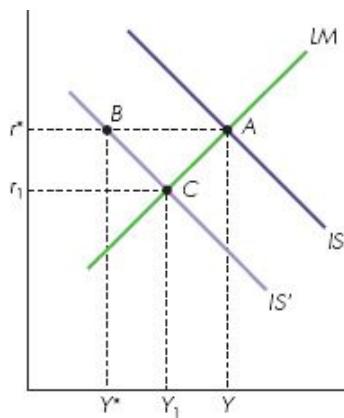


Figure 25.1 Perfect versus imperfect capital mobility

Figure 25.1 illustrates. With a fixed exchange rate and perfect capital mobility, domestic interest rates must match interest rates abroad. The *LM* curve is

horizontal and the supply of money is passively adjusted to maintain this interest rate whatever the level of money demand. Thus, anything that shifts the *IS* curve to *IS'* leads to a fall in short-run equilibrium output as the economy moves from point *A* to point *B*. Of course, this will affect what then happens to domestic prices. We discuss this shortly.

Under perfect capital mobility, vast quantities of financial capital move from assets in one currency to assets in another if there is any perceived difference in expected return. Sterilized intervention does not work because offsetting capital flows are immediately induced.

Figure 25.1 allows us to contrast the worlds of perfect and imperfect capital mobility. Since these refer to the freedom, or otherwise, to change interest rates, we need a diagram relating interest rates to aggregate demand, which is precisely the role of the *IS–LM* diagram introduced in Chapter 20. The downward-sloping *IS* schedule shows, other things equal, how changes in interest rates affect aggregate demand – lower interest rates increase aggregate demand.

Suppose capital mobility is low and that the central bank can pursue whatever domestic monetary policy it chooses, illustrated by the upward-sloping green *LM* schedule. We can interpret this as a flexible inflation target, in which the central bank raises interest rates when output is higher (because this is associated with higher inflation).

Beginning from point *A*, an adverse demand shock (fiscal contraction, fall in export demand, higher saving rate) shifts the *IS* schedule downwards from *IS* to *IS'*. If the interest rate remained r^* , then the demand shock would move the economy from *A* to *B*, leading to a large fall in output. However, with imperfect capital mobility, the central bank cushions some of this fall in output. No longer so worried about inflation, it reduces interest rates and moves down the blue *LM* schedule to reach point *C* instead of point *B*.

Under **perfect capital mobility**, no monetary cushion is available. The interest rate must remain at r^* to defend the fixed exchange rate. Output falls all the way to Y^* . Perfect capital mobility plus a fixed exchange rate target means shocks to aggregate demand lead to large changes in output. Monetary policy is no longer available as a stabilizer.

CASE 25.1

REINTRODUCING CAPITAL CONTROLS?

We need to distinguish carefully between two questions:

- Are capital controls possible?

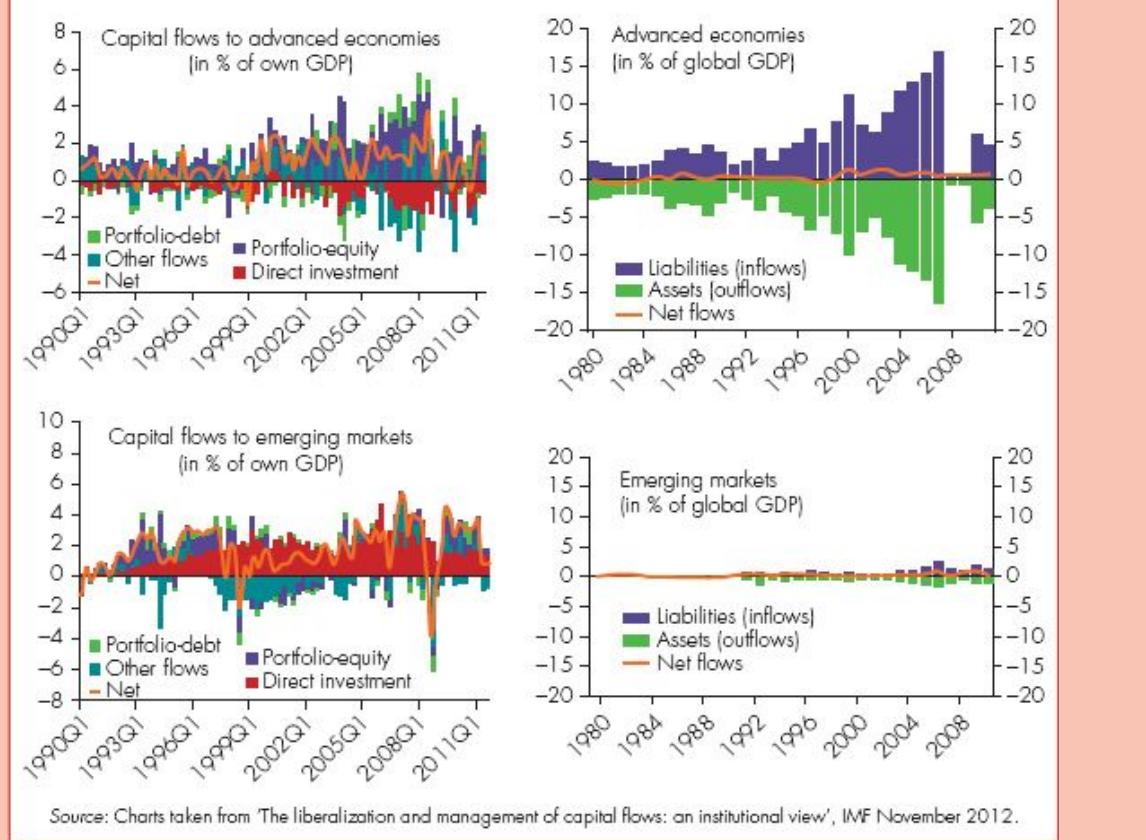
- Are capital controls desirable?

Just as global Internet trading has undermined many national systems of taxation and regulation – for example, national betting tax was made impossible by the presence of Internet gambling sites around the world – so the spread of sophisticated electronic trading in financial assets made it harder to enforce national controls on cross-border financial flows. As technological advance stimulated globalization of financial markets during the 30 years before 2008, people concluded that enforcement of capital controls had become very difficult in countries well-integrated into the global economy.

Some of the less-developed economies were still relatively isolated from the global economy – either as a matter of policy (Myanmar, North Korea) or because their economies were too small and too undeveloped to justify the cost of setting up financial systems to integrate them. For such countries, capital controls remained a possibility.

Suppose capital controls are feasible. Are they desirable? Imagine a group of countries each completely isolated from the asset markets of the rest of the world. This is likely to have several implications. First, within such a fortress economy there is no competition from foreign financial firms, a potential source of inefficiency. Second, unless the country is the United States or China, its small economic size may limit opportunities for scale economies in financial services. Third, without easy access to foreign lending and foreign borrowing, trade in goods and services may be curtailed. This restricts the diversity available to consumers and the scale economies available to producers.

Capital controls come with a heavy price tag – but they offer one offsetting benefit. By removing international asset traders (speculators¹), they make it easier for the domestic government to influence the exchange rate separately from the use of domestic interest rates. With the speculators gone, the government/central bank is the biggest player.



These figures document the steady increase in global capital flows in the decades leading up to the financial crisis, and the reduction in flows subsequently – not yet because policy had changed but simply because investors in rich countries became more sceptical about the profits to be had from investing overseas. There was a 'flight to quality', particularly into bonds issued by the US and German governments, which, as a result, were able to borrow long term at very low rates of interest.

The subsequent financial crash led to reconsideration of the desirability of unfettered capital mobility, for several reasons:

- *Volatility of capital flows* Capital inflows may be an easy way to access funding, but most of this is usually short term. What comes in can easily flow out. If international lenders committed to long-term loans, much of this objection would be dissipated, but most international flows are short term, sometimes called 'hot money'.
- *Economic sovereignty* Adverse capital flows cause huge problems for the government in power. If all lending and borrowing was domestic, the ebb and flow of loans would amount to transfer payments between citizens. The principal effect is on wealth distribution within the country, not on the size of national wealth. When foreign funds exit the country, they can topple government policy itself.
- *Financial contagion* once financial markets become concerned about a country with particular attributes (such as Greece in 2010), they tend to

look around for other countries with some of the same attributes and immediately attack those. Portugal, Italy and Spain are more likely to face a speculative attack if speculators have already succeeded against Greece. If Greece repels the attack, A Tobin tax is a small tax on capital flow transactions. Other countries may escape lightly. Thus, the outcome for each country may depend not only on its own performance but also on the performance of other countries with which it is identified.

In November 2009, UK Prime Minister Gordon Brown resurrected an old idea, the Tobin tax, named after Yale professor James Tobin. Since many financial transactions entail moving huge volumes of money for tiny profit margins, Tobin argued that a tiny ‘transactions tax’ could seriously inhibit international capital flows; throwing sand in the wheels of this activity would restore a greater degree of national autonomy by limiting international capital mobility. For several decades, economists and bankers took the view that the tax might not be desirable in theory and was certainly impossible to implement in practice: smart financial traders would find other untaxed ways of doing the same business. Notice that such a tax would make little difference to the return on an investment held for ten years but a large difference to the return on holding a foreign asset for ten minutes. It would skew capital flows away from short-term hot money.

The motive for resurrecting the proposal was not merely to limit international capital flows but also a desire to limit speculative behaviour in large financial institutions that, being ‘too big to fail’, had de facto insurance from the taxpayer.

In highly connected global financial markets, regulation or taxation by a single country in isolation drives mobile financial business elsewhere. Even if a Tobin tax is desirable – which remains contentious – it would require simultaneous introduction in most important financial centres for it to be effective. The US remains resolutely opposed to the idea. However, the Hollande government in France introduced such a tax in 2012, and the EU Commission is pressing for this to be extended throughout the Eurozone.

A **Tobin tax** is a small tax on capital flow transactions.

IMF chief economist Olivier Blanchard, drawing lessons from the financial crisis, observed that perfect international capital mobility is an idealized benchmark against which to compare actual capital mobility, which would differ in different countries. Where countries have a small margin of manoeuvre, they could use this to pursue expected rates of return on assets that might differ slightly from the world average. But when major strains emerge – as in the case of Greece in 2010 – the extent of national autonomy from speculation is quickly revealed to be small.

In November 2012, the IMF issued a policy paper declaring that capital controls were no longer bad everywhere and always. Judiciously applied, ideally for limited periods rather than in perpetuity, controls could reduce contagion and provide extra policy options.

Adjustment to shocks

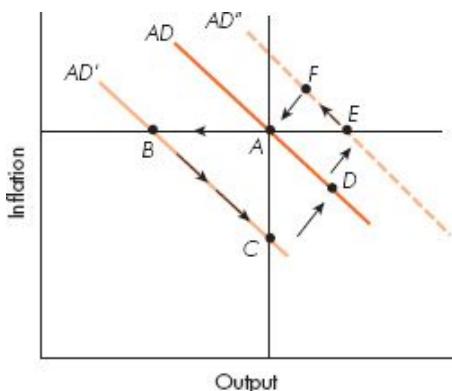
With a fixed exchange rate, how does the economy adjust to a shock when the government takes no monetary or fiscal action to accommodate it? Suppose there is a fall in desired consumption spending at each output level. In a closed economy, output would fall, thus reducing inflation. The central bank would reduce interest rates, thus boosting aggregate demand again. Eventually, internal balance is restored.

What happens in an open economy with a fixed nominal exchange rate and high capital mobility? After the adverse demand shock, there is still a domestic slump. However, any fall in interest rates will generate a massive capital outflow. Interest rates cannot be reduced. Since money demand falls because of lower output, the central bank must reduce the domestic money supply, in line with lower money demand, to *prevent* a change in interest rates.

The adoption of a fixed exchange rate precludes the pursuit of a Taylor rule or inflation target. Interest rate policy has to take care of the exchange rate objective. Any attempt to set interest rates at a different level immediately prompts a massive capital inflow or outflow that changes the money supply and restores the equilibrium interest rate to the required level.

With interest rates thus fixed, the adjustment mechanism of a closed economy is blocked. Lower output and falling prices no longer trigger interest rate cuts that boost aggregate demand again. So how does the economy now get back to long-run equilibrium?

Figure 25.2 illustrates. Suppose initially that both the domestic and the foreign country had 2 per cent inflation. With a fixed nominal exchange rate, the real exchange rate was also constant. A domestic slump then reduces domestic inflation. With a fixed nominal exchange rate, the real exchange rate depreciates, thus raising competitiveness and net exports. An initial fall in domestic absorption ($C + I + G$) eventually induces a sufficient rise in competitiveness to raise net exports ($X - Z$) to restore internal balance. However, higher net exports imply a current account surplus. External balance is not yet restored.



Beginning from internal and external balance at A, a fall in domestic demand shifts aggregate demand from AD to AD' , moving the economy initially to B. The recession bids down inflation, moving the economy down AD' to C. It now has lower inflation than its competitors, and gains competitiveness, eventually shifting aggregate demand back to AD . Although point D restores internal balance, there is a current account surplus because of the gain in competitiveness. With inflation still below that of competitors, aggregate demand keeps rising, taking the economy to E. It then requires a period of aboveaverage inflation, at point F, to eliminate the competition and allow eventual return to internal and external balance at A.

Figure 25.2 Adjustment under fixed exchange rates

Current account surpluses raise the country's net foreign assets. By not spending all its foreign income, the country is saving and getting wealthier. Higher wealth raises consumption demand and domestic absorption. The consequent boom raises inflation, reduces competitiveness at the fixed exchange rate and net exports fall again.

Domestic absorption ($C + I + G$) rises and ($X - Z$) falls. Eventually, both internal and external balance are restored *without any change in interest rates during the adjustment process*. Instead, in an open economy with a fixed exchange rate, adjustment is achieved through temporary booms and slumps that temporarily affect inflation, with induced effects on the real exchange rate, the balance of payments and changes in external wealth.

A shock from abroad

Suppose next there is a shock to foreign demand, raising demand for net exports. The current account ($X - Z$) moves into surplus. Aggregate demand shifts up, and the economy has a boom and a current account surplus. It adds to forex reserves.

In a closed economy, the boom induces a rise in interest rates that eventually returns aggregate demand to potential output. In an open economy with a fixed exchange rate, interest rates remain constant. The boom gradually bids up inflation and reduces competitiveness, reversing the original rise in net exports. When prices rise enough to restore current account balance, aggregate demand reverts to its original level and internal balance is also restored.² Thus a temporary

period of extra inflation permanently raises the price level, permanently changing the real exchange rate.

An open economy *can* return to internal and external balance under a fixed exchange rate, without assistance from interest rates. Otherwise, monetary union (a permanently fixed exchange rate between member states) would be a non-starter!

However, the speed with which internal and external balance are restored depends a lot on the flexibility of wages and prices. The faster inflation adjusts, the faster the real exchange rate changes and the less a recession is needed to accomplish the required change in competitiveness.

The founding fathers of the Eurozone understood that their monetary union would be more successful if their member states had flexible labour markets, so that domestic wage and price adjustment could act as a substitute for exchange rate changes. There was always a tension between the political desire to include the Club Med countries (Portugal, Italy, Greece and Spain) and the economic concern that their rigid and highly unionized labour markets might find it hard to adjust in a crisis.

Can domestic adjustment be assisted by monetary and fiscal policy?

25.2 Macroeconomic policy under fixed exchange rates

Monetary policy

Interest rates are dedicated to defending the exchange rate when capital mobility is perfect. The higher is capital mobility, the less is the scope for an independent domestic monetary policy. If monetary policy cannot speed adjustment back to long-run equilibrium, can fiscal policy do better?

Fiscal policy

Although a fixed exchange rate, plus perfect capital mobility, undermines the scope for monetary policy, it actually enhances the effectiveness of fiscal policy.

In a closed economy, in the short run a fiscal expansion raises output. The central bank responds by raising interest rates, moderating the output increase but helping to stabilize inflation. In an open economy, monetary policy keeps the interest rate fixed. Fiscal expansion no longer crowds out other components of aggregate demand via higher interest rates. Hence, fiscal policy is more powerful under fixed exchange rates than in a closed economy.

Hence, any shocks to domestic demand can be offset by a fiscal policy to help restore internal balance more quickly. If the change in domestic demand was the only reason that the current account departed from external balance, fiscal expansion will also restore external balance.

Fiscal policy is potentially an important policy weapon under fixed exchange rates, especially since monetary policy can no longer be used. Fiscal policy acts in two ways: automatic fiscal stabilizers that dampen the multiplier, and discretionary changes in government spending and tax rates.

Having analysed the economy with a *given* exchange rate, we now analyse *changes* in the pegged exchange rate.

25.3 Devaluation

A devaluation (revaluation) reduces (increases) the par value of the pegged exchange rate.

Even where exchange rates are pegged at fixed values, occasional adjustments in these **par values** sometimes occur.

During three decades after 1945 the major countries agreed to fix their exchange rates, with occasional adjustments or realignments of these par values. Sterling was devalued in 1949 and 1967, before finally floating in 1973. The general idea was to keep exchange rates fixed for long periods, if possible.

The **par value** is the exchange rate that the government agrees to defend.

We distinguish between effects in the short, medium and long run. Initially, we assume that the domestic country begins from internal and external balance. This lets us highlight the effect of the devaluation itself. Then we consider whether devaluation is an appropriate policy response to a shock that has already moved the economy from its long-run equilibrium position.

The short run

When prices and wages adjust slowly, the immediate effect of a devaluation is to reduce the real exchange rate, thus improving the country's competitiveness. Resources are drawn into domestic industries that compete with imports and into export industries that compete in foreign markets.

Although devaluation tends to raise the quantity of net exports ($X - Z$), *the initial response may be slow. Overnight, there are contracts outstanding that were struck*

at the old exchange rate. It also takes time for buyers to adjust to the new prices they face and for sellers to build up production capacity to supply more.

Hence, in the very short run, devaluation may not improve the trade balance – the value of exports minus imports. Suppose we measure the current account in pounds. If domestic prices of export goods are unchanged and the quantity of exports has yet to rise much, export revenues rise only a little in the short run. Import quantities have not yet fallen much. If their foreign prices are unchanged, their price in pounds rises by the amount of the devaluation. Hence, the value of imports in pounds may rise substantially. In *value* terms, the current account initially gets worse.³ However, in the longer run, as quantities adjust, higher export quantities and lower import quantities improve the trade balance.⁴

The medium run

Domestic output Y equals aggregate demand, which is domestic absorption ($C + I + G$) plus net exports

$(X - Z)$. Once quantities begin to adjust, devaluation increases net export demand $(X - Z)$. What happens next depends crucially on aggregate supply.

An economy with Keynesian unemployment has spare resources with which to make extra goods to meet this rise in aggregate demand. But if the economy begins at potential output, it cannot produce many more goods. Higher aggregate demand bids up prices and wages. Competitiveness falls, undoing the gain in competitiveness achieved by devaluation. When domestic prices and wages have risen as much as the exchange rate was initially devalued, the real exchange rate and competitiveness return to their original levels. If the economy began from internal and external balance, long-run equilibrium is now restored.

If devaluation is meant to raise net exports for a sustained period, for example to raise more money to service foreign debts, this is compatible with internal balance [$Y^* = (C \text{ foreign debts, this is compatible with internal balance } I + G) + (X - Z)$] only if domestic absorption ($C + I + G$) is permanently cut, for example by tightening fiscal policy.

Thus, beginning at full employment, devaluation *accompanied* by higher taxes will raise the demand for net exports without increasing total aggregate demand. Since there is no upward pressure on domestic prices, higher competitiveness can be sustained in the medium run.

The long run

Can altering the *nominal* exchange rate permanently change the value of *real* variables? Suppose devaluation is accompanied by tighter fiscal policy to allow the economy to meet the higher demand for net exports without any direct

upward pressure on prices. Although this takes care of demand-side effects on prices, we must also think about supply-side effects.

Domestic firms importing raw materials want to pass on these cost increases in higher prices. Workers buying imported TVs realize that import prices are higher and demand higher nominal wages to maintain their real wages. These price and wage rises lead other firms and other workers to react in similar fashion.

In the absence of any real change in the economy, the eventual effect of a devaluation is a rise in all other nominal wages and prices in line with the higher import prices, leaving all real variables unchanged. Eventually, devaluation has no real effect. Most empirical evidence suggests that the effect of a devaluation is completely offset by a rise in domestic prices and wages after four or five years.

Figure 25.3 summarizes this discussion. A one-off nominal devaluation leads to an instant real devaluation that is gradually unwound again; to a rise in output that is gradually unwound as the real exchange rate stimulus wears off; and to a complicated response in the current account balance in value terms. Initially, the devaluation is effectively a price cut – until quantities can respond, making exports cheaper actually harms export revenue. In the medium run, the induced quantity rise in exports benefits the current account in value terms, provided quantities respond sufficiently to price incentives. Eventually, since the real exchange rate is restored to its original level, so is the current account.

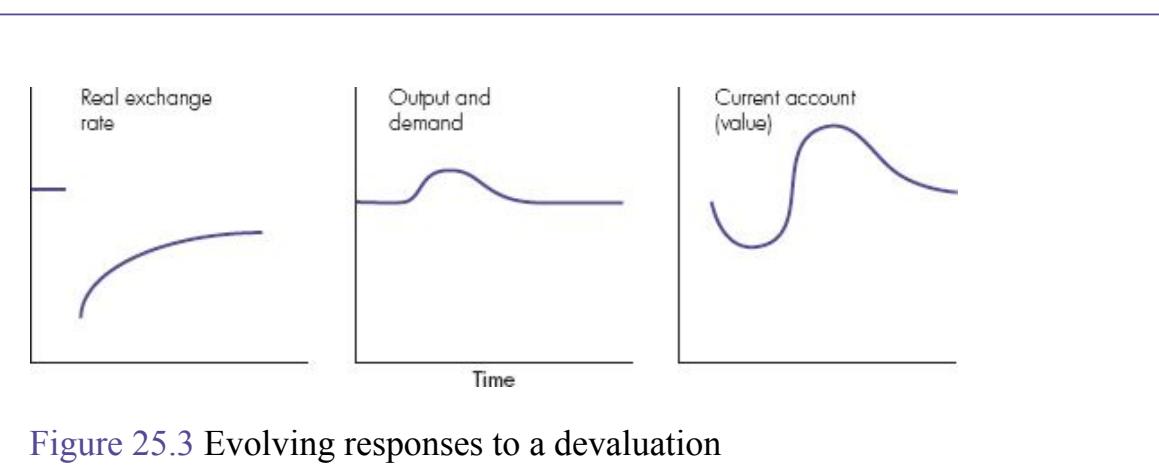


Figure 25.3 Evolving responses to a devaluation

Devaluation and adjustment

To sum up, once quantities begin adjusting, devaluation leads to a temporary but not a permanent rise in competitiveness relative to the path that would have occurred without the devaluation. In the long run, real variables are determined by real forces. Changes in one nominal variable eventually induce offsetting changes in other nominal variables to restore real variables to their equilibrium values.

But devaluation may be the simplest way to change competitiveness *quickly*. It is a useful policy when the alternative adjustment mechanism is a domestic slump and a protracted period of lower inflation until competitiveness is increased.

Suppose there is a permanent fall in export demand. At the original exchange rate, this generates a slump that induces a period of lower inflation, which reduces wages and prices enough to boost competitiveness and restore current account balance. But this takes several years. Devaluation accomplishes an overnight improvement in competitiveness. It speeds up adjustment.

Devaluation may therefore be an appropriate response to a real shock that requires a change in the equilibrium real exchange rate. Conversely, where no real change is required, devaluation eventually generates rises in prices and nominal wages. Chapter 22 discussed inflation expectations and credibility.

Economies can get locked into self-fulfilling prophecies of high inflation. In such circumstances, maintaining a constant real exchange rate requires a steady reduction in the nominal exchange rate. One way to accomplish this is by regular devaluations.

Devaluation has a bad name because it is often associated with periods of high inflation and weak government. This is correct. However, even well-designed macroeconomic policy might choose occasionally to realign the nominal exchange rate. The appropriate circumstance would be a large and sustained shock to the trade balance.

MATHS 25.1

AGGREGATE DEMAND UNDER FIXED EXCHANGE RATES

In previous chapters, we have assumed that: (a) aggregate demand depends negatively on real interest rates and (b) monetary policy raises real interest rates when inflation and output are higher.

$$\begin{aligned} Y &= A - b(r - \pi) & A > 0, b > 0, \\ r - \pi &= f\pi & f > 0 \end{aligned} \tag{1}$$

Putting these together, aggregate demand can be written as:

$$Y = A - a\pi \quad a = bf \tag{2}$$

where A captures all the other influences on aggregate demand (investment expectations, government demand, export demand) and the second term shows that higher inflation π reduces aggregate demand for output by inducing the central bank to raise real interest rates.

Under fixed exchange rates, nominal interest rates are fixed and monetary policy is entirely dedicated to maintaining that level, whatever the

fluctuations in money demand. So what now is the relationship between inflation and aggregate demand once the ii schedule is suppressed?

Aggregate demand is now:

$$Y = A^I - b(r - \pi) - h(EP/P^*) \quad A^I > 0, h > 0 \quad (3)$$

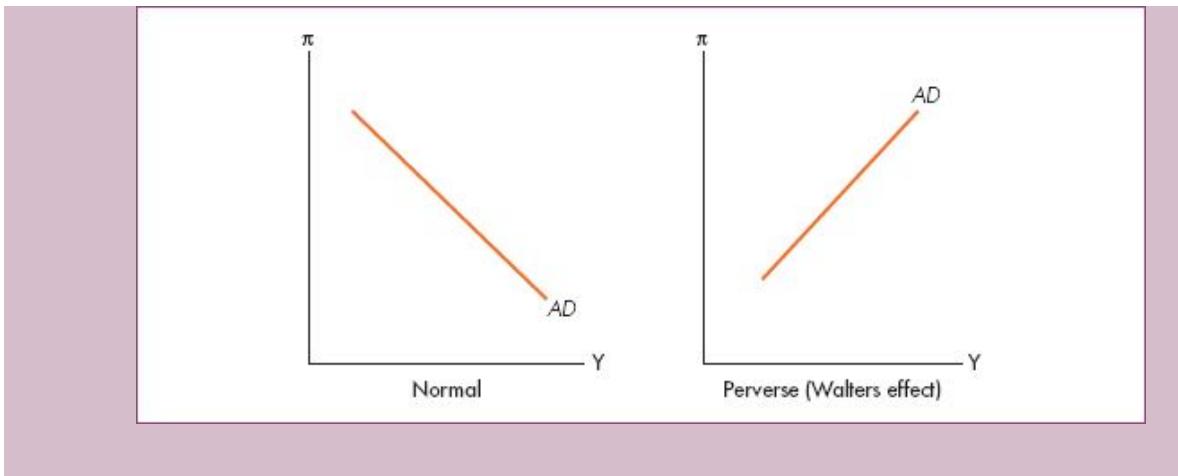
where the third term shows the effect of a higher real exchange rate on demand for domestic output. The nominal exchange rate E is fixed under a pegged exchange rate regime, and we can treat foreign prices P^* as exogenous. We are thus left with two effects.

First, a rise in domestic inflation *increases* domestic aggregate demand by reducing the real interest rate when nominal interest rates must be held pegged to peg the exchange rate. This is sometimes called the ‘Walters effect’ after Margaret Thatcher’s economic advisor Professor Alan Walters. Walters was sceptical of UK membership of any European fixed exchange rate club because of this ‘perverse effect’ – a country with inflation gets a lower real interest rate, boosting aggregate demand and causing yet more inflation and yet more demand, a potentially unstable spiral.

Of course, if things were that simple, the Eurozone would have blown up long ago. The reason it has not blown up is because of the stabilizing effect of the final term in equation (3). With a pegged nominal exchange rate, a country with higher inflation becomes less and less competitive as its real exchange rate appreciates. Lower competitiveness *reduces* aggregate demand.

Provided the second effect is larger than the first, higher inflation reduces aggregate demand. It is therefore still possible to construct a downward-sloping aggregate demand schedule, when plotted against inflation, even when there is a nominal exchange rate peg. Most of the analysis of previous chapters can then be reused.

What circumstances give rise to the ‘normal’ downward-sloping aggregate demand schedule? This will occur when the real exchange rate effect on competitiveness is larger than the real interest rate effect on investment. This is likely to be the case for a small open economy whose trade sector is large. The pressure to create a monetary union in Europe did not arise by accident – it arose because small open economies trading a lot with one another were very sensitive to real exchange rate changes. They wanted to create stability in their competitiveness and to preclude ‘exchange rate wars’ between themselves. This tends to be evidence in favour of the importance of the real exchange rate effect in equation (3). Economists usually use devaluation (revaluation) to describe discrete falls (rises) in pegged exchange rates, but depreciation (appreciation) to describe falls (rises) in floating exchange rates.



25.4 Floating exchange rates

Having discussed fixed exchange rates, we now turn to the opposite case, freely floating exchange rates. The foreign exchange reserves remain constant, the balance of payments is zero and the government refrains from any intervention in the forex market. In this section we explain how the level of the exchange rate is determined in the short run. The next section uses this analysis to study monetary and fiscal policy in an open economy with floating exchange rates.

The long run

In long-run equilibrium the economy is at both internal and external balance. Chapter 24 analysed determinants of the real exchange rate in long-run equilibrium. Given that output is at potential output, this real exchange rate must achieve current account balance. Anything that tends to create a current account surplus (a resource discovery, a new export industry, income from foreign assets) induces a real exchange rate appreciation to reduce competitiveness. This reduces net exports until external balance is restored.⁵

When exchange rates float freely, there is no official intervention in the forex market and no net monetary transfer between countries since the balance of payments is always zero. Just as in a closed economy, the central bank controls the domestic money supply or sets it to achieve the interest rate it wishes.

The monetary rule, and associated nominal anchor, then determines the domestic price level as explained in Chapter 22. For example, we can think of inflation targeting as pursuing a target path for the price level. In the long run, a Taylor rule has the same result since interest rates are adjusted until output is restored to potential output and inflation is restored to target inflation.

With perfect capital mobility, the central bank can use interest rates to peg the exchange rate, thereby giving up the independent use of interest rates to

manipulate the domestic economy, or can set interest rates to manipulate the domestic economy but must then accept the level to which the exchange rate floats.

Our theory of floating exchange rates in the long run is thus easily summarized. Real forces determine the long-run equilibrium *real* exchange rate necessary for external balance. Domestic, sovereign, monetary policy determines the path of the domestic price level. Given the path of foreign prices, there is then only one path of the *nominal* exchange rate that achieves the appropriate real exchange rate in the long run.

If domestic and foreign monetary policies generate the same inflation rates, a constant real exchange rate in the long run is compatible with a constant nominal exchange rate in the long run. However, if domestic and foreign inflation rates differ permanently, the nominal exchange rate must change steadily to keep the real exchange rate at its equilibrium level.

Two examples may help reinforce this argument. Suppose first that there is no inflation anywhere. A once-and-for-all change in domestic monetary policy leads to a doubling of the domestic price level. Thereafter prices are constant. To maintain the real exchange rate, there is a once-and-for-all depreciation of the nominal exchange rate by 50 per cent, say from \$2/£ to \$1/£.

Although domestic prices have doubled, the dollar price of UK exports is unaffected in the long run. A £10 shirt used to sell for \$20 at an exchange rate of \$2/£. Now it costs £20 to make the shirt but it still sells for \$20 since the exchange rate is now \$1/£. Similarly, a US baseball bat costing \$40 used to sell for £20 in the UK. After the exchange rate falls to \$1/£, it still costs \$40 to make but now sells for £40 in the UK. The price of UK imports doubles because the exchange rate falls by 50 per cent. Import prices rise in line with domestic prices in the UK. Whether we compare the relative prices of UK and US goods in dollars or in pounds, their relative price is unaltered. Competitiveness does not change.

How about continuous inflation? Suppose US inflation is zero but annual UK inflation is permanently 10 per cent. A steady depreciation of the \$/£ exchange rate, by 10 per cent a year, leaves the real exchange rate and competitiveness constant. The dollar price of UK goods is constant, like US prices, and the pound price of imports from the US rises annually at 10 per cent, just like UK goods.

A **Purchasing power parity (PPP)** is the path of the nominal exchange rate that offsets differential inflation rates across countries, maintaining a constant real exchange rate.

Hence, in the long run, floating exchange rates adjust to achieve the unique real exchange rate compatible with internal and external balance. Knowing monetary policy and the price level, we know the required path for the nominal exchange

rate. In the absence of real shocks, the nominal exchange rate follows the [purchasing power parity \(PPP\)](#) path to achieve the equilibrium real exchange rate. Any real shocks that are not accommodated by changes in monetary policy and the price level will eventually induce changes in the nominal exchange rate to achieve the required change in the real exchange rate.

However, in the short run, the real exchange rate can fluctuate a lot. The stock of internationally mobile funds is now vast. If those funds were all to move in a short period, say an hour, between two currencies, this massive flow on the financial account could not possibly be offset by the small net flows that occur on the current account during that hour. Under freely floating exchange rates there is no government intervention and no official financing. The forex market could not clear.

But clear it does, hour by hour, and indeed minute by minute. Short-run equilibrium in the forex market is achieved because the exchange rate is capable of jumping at any instant to the level necessary to *prevent* one-way capital flows of large magnitude. To examine this process in more detail, we need to think more about capital flows.

The short run

When international capital mobility is perfect, interest parity must hold. If assets in different currencies offer different expected returns, there will be massive one-way traffic in capital flows, which is inconsistent with forex market equilibrium. Hence, expected returns are equated to *prevent* massive one-way capital flows.

Expected returns include capital gains and losses as exchange rates change while foreign assets are held temporarily. The current level of the exchange rate affects this capital gain between now and the next period. Hence floating exchange rates can always be set at a level that makes expected capital gains just offset interest rate differentials across currencies, for example between UK bonds in sterling and US bonds in dollars, as interest parity requires.

Suppose UK interest rates are 2 per cent higher than US interest rates. Why do holders of funds not move all their funds into sterling? If speculators expect the pound to depreciate by 2 per cent a year against the dollar, investors in pounds get 2 per cent extra interest but lose 2 per cent a year on the exchange rate, relative to the alternative strategy of lending in dollars. The extra interest just compensates for the expected loss and most speculators will not mind where they hold their funds. Without massive flows between currencies, the forex market can be in equilibrium.

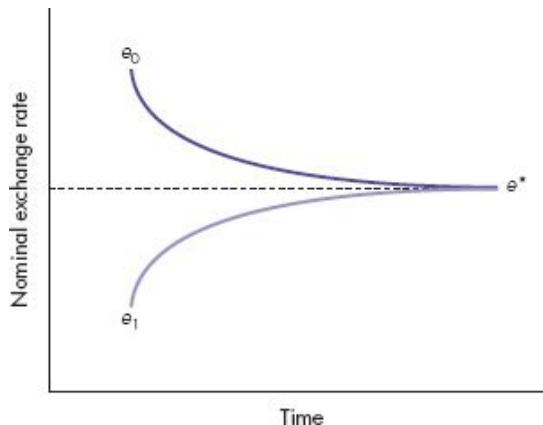
What happens if UK interest rates rise and are now 4 per cent above US interest rates? If people still think that the UK exchange rate will fall at 2 per cent a year, the extra UK interest rate more than compensates for capital losses on sterling. Everyone tries to move into pounds. Almost instantly, this bids up the \$/£

exchange rate. By how much? Until it reaches such a high level that people expect the pound then to fall by 4 per cent a year thereafter. Only then are the capital losses expected on funds lent in pounds sufficient to offset the 4 per cent interest differential. This restores interest parity and ends one-way traffic.

Why does a higher value of the pound today make people expect the UK exchange rate to fall in the future? Because smart speculators figure out that eventually the exchange rate has to return to the level that achieves external balance long-run equilibrium. With the end point anchored, a higher initial value means a faster rate of subsequent fall.

Floating exchange rates are volatile because they are asset prices that reflect beliefs about the entire future. Such beliefs can change a lot.

Figure 25.4 sums up our theory of floating exchange rate determination. For simplicity, suppose domestic and foreign monetary policies generate similar eventual inflation rates. The constant equilibrium real exchange rate in the long run then implies a constant nominal exchange rate e^* to achieve external balance. In the short run, the nominal exchange rate can depart a lot from e^* and can change rapidly.



Suppose e^* is the equilibrium nominal exchange rate in the long run. To be constant, this requires that there is no inflation differential across countries. The Fisher hypothesis then means that interest differentials are also eliminated eventually. For a country with temporarily high interest rates, the exchange rate begins at e_0 and moves along $e_0 \rightarrow e^*$. For a country with temporarily low interest rates, the exchange rate begins at e_1 and moves along $e_1 \rightarrow e^*$. In either case, expected exchange rate changes offset interest differentials.

Figure 25.4 Floating exchange rates

A country with high (low) interest rates in the short run must have a currency expected to depreciate (appreciate) if it is not to generate one-way capital flows in or out of the currency. Figure 25.4 shows two possible paths for the nominal exchange rate. If interest rates are expected to be high by world standards, the

exchange rate begins at e_0 and depreciates steadily until long-run equilibrium is reached at e^* .

At every point along $e_0 e^*$, the slope of this schedule reflects the interest differential, and the capital loss is just offsetting the interest differential.

In the long run, the nominal exchange rate can be constant only if the interest rate differential is eliminated. Recall from Chapter 22 the Fisher hypothesis, which says that nominal interest rates largely adjust in line with inflation, since real interest rates are fairly constant. Figure 25.4 assumes the inflation differential is eventually zero, so that the nominal exchange rate can be constant. This is quite consistent with assuming that eventually monetary policies converge and the interest rate differential disappears.⁶

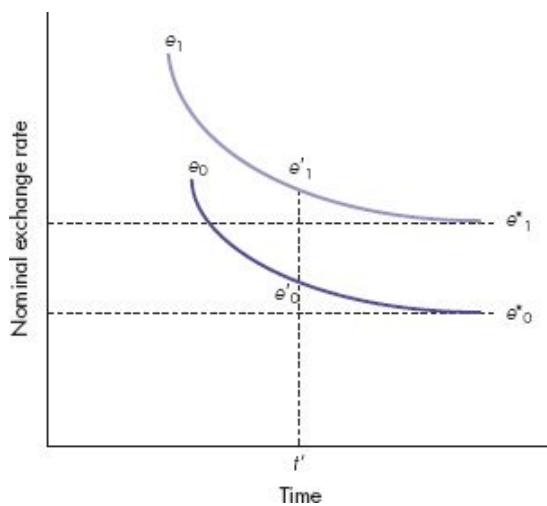
Conversely, if a country is expected to have abnormally low interest rates for a while, its exchange rate will begin at e_1 in Figure 25.4. The foreseen exchange rate appreciation along $e_1 e^*$ provides expected capital gains on the exchange rate to compensate for the low interest rate on sterling-denominated assets.

From time to time, financial markets change their view about the likely future path the economy will follow. This can lead to a dramatic reassessment of the current exchange rate needed to prevent massive capital flows.

Suppose financial markets revise their expectations about the future of interest differentials. Formerly, they believed a country would have high interest rates for a while; now they believe that interest rates will be low by world standards. In terms of Figure 25.4, the appropriate path switches from $e_0 e^*$ to $e_1 e^*$ and the initial exchange rate therefore jumps from e_0 to e_1 .

Any jump in an asset price is unexpected. If people had expected the exchange rate to jump from e_0 to e_1 , they would already have moved out of sterling assets. People holding an asset when its price jumps up or down are either lucky or unlucky.

One reason exchange rates may jump is because of new information about the future of interest rates. Another reason is because of new information about the long-run equilibrium exchange rate. Figure 25.5 shows a country with temporarily high interest rates that was initially expected to have a long-run equilibrium exchange rate e_0^* . Its exchange rate was expected to move along $e_0 e_0^*$ as time elapsed.



While a country has higher interest rates than its partners, its exchange rate must be depreciating, for example along e_0 to e_0^* . Suppose at time t' people first realize the long-run exchange rate will instead be e_1^* . At t' the exchange rate jumps immediately from e_0 to e_1 in order to follow the path from e_1 to e_1^* .

Figure 25.5 A change in the expected long-run exchange rate

At time t' the financial markets get information that the long-run equilibrium exchange rate will in fact be e_1^* . Had they known this all along, the exchange rate would have begun at e_1 and moved along e_1^* . When this is first realized at time t_1 , the forex market immediately jumps the exchange rate to e_1^* so that the exchange rate is appropriate from now on.

Along the path e_1^* the expected capital loss on the exchange rate continues to offset the interest differential. And the path is expected to end up in the right place; namely, at the long-run equilibrium exchange rate.

25.5 Monetary and fiscal policy under floating exchange rates

In a closed economy with slow wage and price adjustment, changes in monetary and fiscal policy have real effects in the short run, although the economy eventually returns to internal balance. In an open economy with *fixed* exchange rates, almost perfect capital mobility makes monetary policy almost powerless in the short run; however, the power of fiscal policy is enhanced since fiscal expansion no longer bids up interest rates. Under *floating* exchange rates the converse is true: monetary policy is powerful in the short run, but the effectiveness of fiscal policy is reduced.

Monetary policy

Figure 25.4 has already displayed the power of monetary policy to affect the real economy in the short run under floating exchange rates. Given the exchange rate expected in the long run, the anticipation of higher interest rates in the short to medium run causes an immediate appreciation of the exchange rate so it is then likely to fall thereafter. Anticipated capital losses from now on are what choke off the capital inflow that high interest rates would otherwise cause.

Conversely, the anticipation of a period of low interest rates (relative to trading partner countries) induces an initial depreciation of the exchange rate, so that it is likely to rise thereafter. The prospect of future capital gains prevents a capital outflow when interest rates are low.

Hence, beliefs about current and future monetary policy can have a dramatic effect on the initial level of the exchange rate and competitiveness. In effect, the exchange rate is pricing beliefs about the entire future of monetary policy, both at home and abroad. Changing the current interest rate for a short time will have only a small effect on this calculation. However, a credible change in monetary policy for a sustained period will cause a large re-evaluation of the correct path for the exchange rate. This can have a large effect in the short run.

Thus in an open economy with floating exchange rates, monetary policy affects aggregate demand not merely through the effect of interest rates on consumption and investment demand. Changing the anticipated path of interest rates can have a large effect on the exchange rate and competitiveness. This effect on aggregate demand may be large. Because the effect of interest rates on competitiveness operates in the same direction as the domestic effect – lower interest rates boost domestic spending, but also induce a lower exchange rate and greater competitiveness, boosting net exports – monetary policy is more powerful under floating exchange rates than in a closed economy.

Fiscal policy

Under floating exchange rates, this effect of interest rate changes on competitiveness *reinforces* the power of monetary policy, but *undermines* the power of fiscal policy.

Suppose the government undertakes a fiscal expansion, raising government spending. This increases aggregate demand. Whether monetary policy follows an inflation target, a Taylor rule or a nominal money target, the boom induces the central bank to raise interest rates. The higher interest rate induces an immediate appreciation of the nominal exchange rate to choke off a capital inflow: if the exchange rate is high enough, people will believe it will fall from now on.

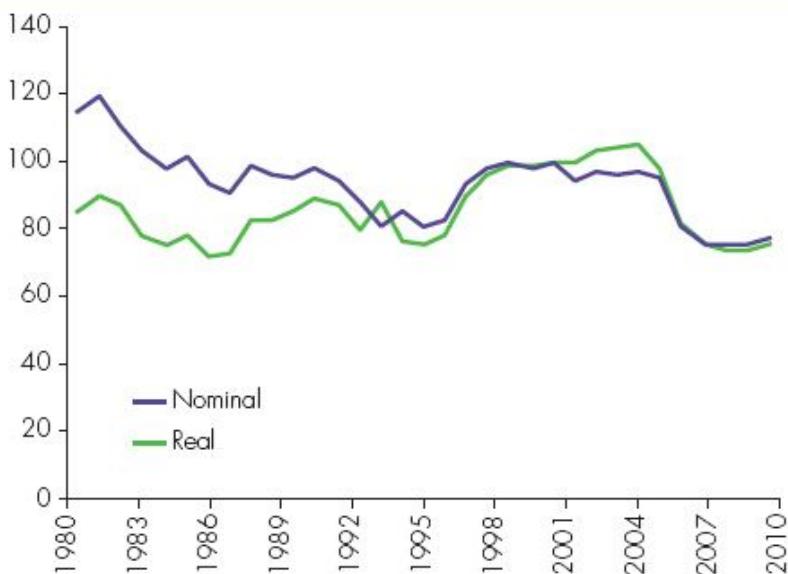
In a closed economy, higher interest rates partially crowd out private expenditure by reducing consumption and investment demand. But in an open economy with

floating exchange rates, the induced exchange rate appreciation also reduces competitiveness and the demand for net exports, further dampening the power of fiscal expansion to stimulate aggregate demand in the short run.

ACTIVITY 25.1

A 30-YEAR LOOK AT STERLING

The figure shows the nominal and real sterling exchange rates since 1980. We show the nominal effective exchange rate (eer) against a basket of the currencies most important for the UK's international trade, weighted by their importance in UK trade. We also show the real effective exchange rate (rer), adjusting the nominal exchange rate for movements of relative prices at home and abroad (using the same weights as used to construct the nominal effective exchange rate).



UK nominal and real exchange rate, 1980–2012 (2002 = 100)

Sources: Bank of England; OECD, *Economic Outlook, Statistical Annex*.

Notice the high correlation between movements in the nominal exchange rate and real exchange rate. In the short run, most changes in the real exchange rate are caused by changes in the nominal exchange rate, not by changes in domestic and foreign prices.

The nominal exchange rate is more volatile than the price of goods or the price of labour. The correlation between nominal and real exchange rates got much stronger after the mid-1990s, when all the major countries adopted inflation targeting with similar inflation targets, largely eliminating differential price trends as sources of real exchange rate movements. We can use this figure to discuss UK exchange rate behaviour during the last 30 years.

In 1980 a tight monetary policy had been introduced to fight inflation. The prospect of high interest rates for some time had led to a sharp appreciation in the nominal exchange rate, precisely so that it fell thereafter, thus generating expected capital losses on holding sterling that would offset the high UK interest rates and prevent a one-way flow of financial capital into the UK. At the same time, the UK had found oil and the rise in oil prices made this oil more valuable. This raises the long-run real exchange rate. Figure 25.5 shows that, in anticipation of such a rise, capital flows begin. Hence the exchange rate appreciates as soon as the discovery is made.

Together, tight money and North Sea oil explain why the real value of sterling was high in 1980. Competitiveness was therefore low and exports of non-oil products, particularly manufactured goods, were badly hit. After 1981, competitiveness improved for a while as the real exchange rate depreciated. By 1988 the real exchange rate index had depreciated by about 12 per cent, from 85 to 75.

By the late 1980s, Chancellor of the Exchequer Nigel Lawson was 'shadowing' the monetary policy of the likely Eurozone countries, partly to see if the UK would be suitable as a potential Eurozone member. As it happened, German interest rates were quite low at the time; too low for what was good for the UK. The ensuing boom caused the UK economy to overheat. As inflation took off again, the UK was forced to raise interest rates sharply in the late 1980s. As you would expect, the figure shows that this induced a new exchange rate appreciation, of both nominal and real exchange rates, after 1987.

In 1990 the UK joined the Exchange Rate Mechanism – the prelude to the Eurozone – and pegged its nominal exchange rate to other EU countries.⁷ To sustain the peg, the UK had to match interest rates in other ERM countries. The timing was lousy: 1990 was the year of German reunification. Soon Germany was giving big budget subsidies to support East Germans until their productivity caught up to West German levels. Given this fiscal expansion in Germany, it took very high interest rates in the ERM to restrain inflation in Germany.

Like many other ERM members, the UK was crippled trying to match these huge interest rates that were fine medicine for Germany but not for its partners. By 1992 the UK problem was no longer overheating but a deepening recession caused by very tight monetary policy. Since most of the UK's trade partners also had high interest rates, the pound did not appreciate much during 1990–92.

In September 1992 the UK left the ERM, floated the exchange rate, and announced that it was cutting interest rates to end recession. The figure shows the consequent and substantial depreciation of nominal and real UK

exchange rates in late 1992. Unlike the UK, most ERM members stayed in the system and staggered on under high interest rates for several years more.

Greater competitiveness gave the UK an export boom during 1993–95 that helped pull it out of recession. However, the figure shows that by 1997 the real exchange rate was back above its level of 1992 when the UK had left the ERM. Sterling continued to appreciate thereafter. Most EU countries were now tightening fiscal policy to meet the Maastricht criteria for monetary union. With tighter fiscal policy, their monetary policy no longer had to be so tight. As their exchange rates depreciated, the pound appreciated.

In addition, Labour looked likely to be the party of government for several years to come. Independence for the Bank of England guaranteed that interest rates would rise if required to keep inflation under control. Despite the Code for Fiscal Stability and the Chancellor's emphasis on prudence, financial markets were never convinced about Labour's commitment to tight fiscal policy. Markets were already wondering when higher future interest rates would become necessary.

UK exchange rates are not of course the result only of UK policy. In particular, the data in the figure also reflect a major weakening of the US dollar after 2000: a low dollar means a high pound and a high euro. Global investors woke up to the fact that the US had been living beyond its means. It took a depreciation of the dollar to begin to restore US competitiveness. One side effect was the rise in the pound.

Although the government continued to describe the UK economy as sustainable – ‘you've never had it so prudent’ – the figure shows just how much sterling had appreciated in real terms during 1996–2006, prior to the financial crash. This helps explain why the UK current account was in substantial deficit – financial services apart, most of the UK economy had become quite uncompetitive at the high real exchange rate, which could be sustained only as long as global investors remained happy to pile into London, creating capital inflows on the financial account to offset the current account deficit.

The financial crisis pricked international confidence in the UK economy, which was unusually exposed to the financial sector. As capital inflows abated, it became evident that a lower real exchange rate was needed in order to boost the international competitiveness of other sectors of the UK economy. The ensuing real depreciation, by over 25 per cent, is the largest change in competitiveness experienced by the UK economy during the last 30 years.

What should we expect its effect to be? First, increased competitiveness will have a significant effect on the trade balance – higher exports and lower imports – but this will take one to two years to feed through. Depreciation

takes time to affect behaviour. The size of this stimulus will also depend on the growth rate of UK export markets. If export markets are weak, competitors may cut prices to defend their market share, thus reducing the impact of the exchange rate change; if export markets are strong, the stimulus is likely to be larger and be experienced more quickly. In this context, it is helpful to the UK that China and India are booming again, but very unhelpful that its largest trading partner, the Eurozone, is still stagnating.

Following the reduction in aggregate demand, the UK has spare capacity. Hence, there is little danger of an increase in exports crowding out other categories of aggregate demand.

To the extent that higher competitiveness implies output and income are higher than they would otherwise have been, tax revenue will also be higher than it would otherwise have been. Thus, following the financial crash and considerable increase in government indebtedness, the real depreciation of sterling is welcome from a fiscal viewpoint as well as an employment viewpoint.

Questions

- a. Why are there considerable differences between the evolution of nominal and real exchange rates before 1996 but such a close correlation thereafter?
- b. The real exchange rate was systematically and substantially higher during 1998–2006 than previously. If markets had concluded that a higher sterling real exchange rate was permanently sustainable, what underlying factors could have led to such a conclusion?
- c. Suppose the UK had been a member of the Eurozone after 2000. How would the response of the UK economy to the financial crisis have differed? Would the UK have experienced more or fewer problems? What would have been the alternative to an exchange rate depreciation?

To check your answers to these questions, go to page 681.

Summary

- With a **fixed exchange rate** and **perfect capital mobility**, the domestic interest rate must match foreign interest rates to prevent massive capital flows and allow equilibrium in the forex market. **Monetary sovereignty is then lost.** Monetary policy cannot be used independently to control the domestic economy.

- A **fall in domestic demand** causes a fall in output and a decline in prices. Unlike a closed economy, monetary policy cannot respond by cutting interest rates. Rather, the fall in prices boosts competitiveness and raises aggregate demand. When internal balance is restored, there is now a current account surplus. This generates greater wealth, thus raising domestic demand again. After a temporary boom to raise prices and reduce excess competitiveness, internal and external balance can be restored.
- A **fall in export demand** generates a slump. This lowers prices and increases competitiveness restoring internal and external balance. No subsequent boom is then required.
- In the short run, fiscal policy is a powerful tool under fixed exchange rates. Fiscal expansion no longer bids up domestic interest rates in the short run. Output expansion is accompanied by a rise in the money supply to maintain interest rates at the world level.
- A **devaluation** is a fall in the value of the fixed exchange rate. With sluggish price adjustment, it raises competitiveness and aggregate demand. With spare resources, output rises. But at potential output, net exports can rise only if domestic absorption is cut by tighter fiscal policy.
- In the long run, devaluation is unlikely to have much effect. Changing one nominal variable merely leads to offsetting changes in other nominal variables. In passing on higher import prices and seeking cost-of-living wage increases, firms and workers offset the competitive advantage of devaluation. But devaluation may speed up adjustment to a shock that requires a permanent change in competitiveness to restore internal and external balance.
- Under **floating exchange rates**, the long-run level of the nominal exchange rate achieves external balance, given prices at home and abroad. In the short run, the exchange rate adjusts to prevent massive flows on the capital account.
- The exchange rate must begin at a level from which the anticipated convergent path to its long-run equilibrium continuously provides capital gains or losses to offset expected interest rate differentials, thus equating the expected return on lending at home and abroad.

- Under floating exchange rates, monetary policy is a powerful short-term tool. The belief that interest rates will be higher for some time induces a sharp appreciation of the exchange rate, so that it can then credibly promise capital losses to offset high interest rates. With sluggish price adjustment, the initial appreciation of the nominal exchange rate causes a sharp fall in competitiveness. This reduction in demand for net exports reinforces other effects of high interest rates in reducing aggregate demand.
- Fiscal policy is a weaker tool under floating exchange rates. Fiscal expansion induces a boom and higher interest rates. The latter induce an exchange rate appreciation that crowds out some net exports, reinforcing domestic crowding out of consumption and investment.
- The actual path of the UK nominal exchange rate reflects changing beliefs about the future course of domestic and foreign interest rates, and about the eventual level of the exchange rate in long-run equilibrium. The latter depends on beliefs about the eventual price level at home and abroad, but also on supply shocks such as resource discoveries.

Review questions



EASY

- 1 Which of the following statements is correct? Devaluation is most effective when: (a) a country has a small export and import sector, since higher import prices then have little effect; (b) domestic wages and prices are very flexible; (c) nominal wages and prices adjust slowly; (d) the country is already at potential output.
- 2 Beginning at internal and external balance, an economy devalues its nominal exchange rate by 30 per cent. (a) What is the eventual change in the price level, nominal wages and the nominal money supply? (b) What is the nominal anchor in this economy?
- 3 A country discovers a new technology that will add significantly to its export capacity in five years' time. (a) What must happen to its real exchange rate in the long run? (b) Why does the exchange rate react immediately to the news rather than waiting till the new export supply comes on stream?
- 4 **Common fallacies** Why are these statements wrong? (a) Collectively, global speculators have more money than central banks. Hence, central banks cannot defend fixed exchange rates. (b) Floating exchange rates are volatile because imports and exports fluctuate a lot. (c) Exchange rate policy is really monetary policy, so it makes no difference to the impact of fiscal policy.

5 ‘Once the central bank is made independent, with a specified inflation target, the principal role of macroeconomic policy is to determine the real interest rate and hence the exchange rate.’ Explain.

MEDIUM

- 6 Because of China’s sustained export success, many people in the West call for China’s fixed exchange rate against the dollar to be revalued or for its currency to be floated in the expectation that it will then appreciate. (a) At its current stage of development, should China be running a deficit or surplus on the financial account of its balance of payments? (b) Given that its trade surplus in 2006 exceeded \$170 billion, was China running a balance of payments surplus or deficit? (c) With such large monetary inflows, what was happening to China’s foreign exchange reserves and the Chinese money supply? Must this be inflationary, or could the demand for money increase just as quickly?
- 7 Manchester has a local government but a fixed exchange rate with the rest of the UK. How powerful are the monetary and fiscal policies of the local government in Manchester?
- 8 Beginning at internal and external balance, an economy devalues its fixed exchange rate. (a) What happens to its real interest rate in the long run? (b) What happens to output? (c) What happens to inflation? (d) How are internal and external balance restored?
- 9 A country faces a permanent fall in export demand. Would devaluation help? How else might internal and external balance be restored?
- 10 Because of the strength of long-run Asian demand for its mineral exports, markets conclude that the Australian real exchange rate will have to be permanently higher. Australian monetary policy is already much tighter. Draw a diagram showing the likely evolution over time of the exchange rate of the Australian dollar against sterling.
- 11 Rank the following three situations according to the ability of monetary policy to affect real output in the short run: (a) a closed economy; (b) an open economy with fixed exchange rates; (c) an open economy with floating exchange rates. Explain.
- 12 **Essay question** What do you see as the relative advantages and disadvantages of fixed and floating exchange rates?
- 13 Because of the strength of long-run Asian demand for its mineral exports, markets conclude that the Australian real exchange rate will have to be permanently higher. Australian monetary policy is already much tighter. Suppose Australia now discovers vast new mineral deposits that will take five years to begin to exploit. What further effect, if any, will this have on the evolution of Australia’s exchange rate? Illustrate in a diagram.

HARD

14 The following table shows the evolution of an index of \$/£ nominal exchange rate, and the behaviour of prices in each of two countries. In the initial years, monetary policy is very different; in the last three years both countries succeed in achieving inflation targeting at a low level. For simplicity, we assume that the inflation target is zero. (a) Calculate the evolution of the implied real exchange rate, setting the index initially at 100. (b) Graph the nominal and real exchange rates. (c) What happens to the correlation between nominal and real exchange rates once inflation convergence is achieved?

Period	Different inflation			Inflation convergence		
	1	2	3	4	5	6
Nominal \$/£ index (period 1 5 100)	100	110	90	80	70	90
UK price index (period 1 5 100)	100	110	120	120	120	120
US price index (period 15 100)	100	100	90	90	90	90
Real \$/£ index (period 1 5 100)						

- 1 A speculator is someone who decides to buy an asset, for subsequent resale, believing the expected capital gain will justify the investment. Since almost all assets entail some potential capital gain, there is a fine line between the terms investor and speculator. In popular parlance, the term speculator also presumes that the time horizon before resale is short, and that the activity is likely to be socially damaging.
- 2 During the boom, the current account surplus adds to foreign assets, which may therefore be a little higher in the new equilibrium than in the original equilibrium. If so, restoring current account balance does not quite restore the original level of the trade balance. For internal balance, potential output equals domestic absorption plus net exports. If net exports have changed a little, so has domestic absorption. Such details belong in a more advanced course. The basic adjustment mechanism remains as described in the text.
- 3 The famous Marshall–Lerner condition says that devaluation improves the trade balance only if the sum of the price elasticities of demand for imports and exports is more negative than -1 . Recall from Chapter 4 that, when demand is elastic, the revenue effect of changes in quantity more than offsets the effect of a change in price. In the short run, when demand is inelastic, devaluation may worsen the current account.
- 4 Thus a devaluation first worsens then improves the trade balance, a response known as the *J-curve*. As time elapses after the devaluation, the trade balance falls down to the bottom of the J but then rises above its initial position.
- 5 Economists usually use devaluation (revaluation) to describe discrete falls (rises) in pegged exchange rates, but depreciation (appreciation) to describe falls (rises) in floating exchange rates.
- 6 A different assumption about monetary policy would imply a permanently changing nominal exchange rate.
- 7 A different assumption about monetary policy would imply a permanently changing nominal exchange rate.

CHAPTER 26

Exchange rate regimes

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 contrast different exchange rate regimes
- 2 describe the gold standard
- 3 discuss an adjustable peg
- 4 explain the impossible triad
- 5 examine speculative attacks
- 6 analyse floating exchange rates
- 7 explain the Exchange Rate Mechanism and the Maastricht criteria
- 8 discuss how the Eurozone operates

In a closed economy, money is the medium of exchange that reduces transaction costs. Similarly, a group of nation states using different currencies requires an international means of facilitating cross-border transactions efficiently via an **international monetary system**. Exchange rates convert one currency into another. We review different **exchange rate regimes**, then analyse their relative merits. Finally, we examine whether interdependence of nation states creates a motive to co-ordinate national economic policies.

Table 26.1 identifies four regimes – the gold standard, an adjustable peg, managed float and free float – according to the intervention obligations on the central bank. Any exchange rate involves two countries. Most regimes require that the two governments agree on which regime is in force.

The **international monetary system** provides a medium of exchange for international transactions.

An **exchange rate regime** is a policy rule for intervening (or not) in the forex market.

Table 26.1 Exchange rate regimes

Forex intervention	Exchange rate	
	Fixed	Flexible
None	–	Free float
Automatic	Gold standard	–
Some discretion	Adjustable peg	Managed float

26.1 The gold standard

The gold standard was in force for most of the nineteenth century, though some countries like the UK had adopted it much earlier. It had three key rules.

Each government fixed the price of gold in its domestic currency; gold was convertible domestic currency, at this fixed price, in whatever quantities people wanted to transact; and domestic money creation was linked to the government's stock of gold. Each unit of currency in circulation was backed by an equivalent value of gold in the vault of the central bank. Cash could not be created unless the central bank could acquire gold.

The US gold price was \$20.67 an ounce, the UK A monetary union of different countries is a commitment to permanently fixed exchange rates. price £4.25 an ounce. The \$/£ exchange rate was thus fixed at \$4.86/£, \$20.67 divided by £4.25. At any other exchange rate, people could sell gold in one country and buy gold in the other, making a profit with certainty. The equilibrium exchange rate was the relative gold price in the two currencies.

A **monetary union** of different countries is a commitment to permanently fixed exchange rates.

The gold standard was a **monetary union** based on fixed gold prices, convertible currencies and complete gold backing for the money supply. Because monetary union is a live issue today, it is interesting to ask how the gold standard worked.

Balance of payments adjustment

In long-run equilibrium, each country has internal and external balance. Each country also has a constant money supply, a given level of gold in the central bank vault, a given price level and a constant interest rate.

Suppose Americans now spend more on imports from the UK. The UK has a trade surplus. If domestic prices and wages are slow to adjust, the UK has an export-led boom in the short run. Aggregate demand for UK output rises. Conversely, the US has a recession and a trade deficit.

This provides an *automatic* adjustment mechanism. Initially, the UK has a balance of payments surplus, and hence a rise in foreign exchange reserves (gold). A UK trade surplus provides more gold at the Bank of England, and a matching increase in the domestic money supply. This augments the UK boom. As prices rise, the UK becomes less competitive: the nominal exchange rate remains fixed but the real exchange rate appreciates, eliminating the trade surplus, eventually restoring external balance.

In the US, with a payments deficit, US gold and money falls, raising US interest rates and reducing aggregate demand further. Gradually US prices and wages fall, competitiveness rises, the trade deficit is gradually eliminated and external balance restored.

The gold standard provided an automatic mechanism for adjusting imbalances in trade and payments. However, adjustment was slow. Since it depended on changes in domestic wages and prices to adjust competitiveness, the speed of adjustment reflected the speed with which domestic prices and wages adjusted to excess supply or excess demand.

CONCEPT 26.1

THE GOLD STANDARD AND CAPITAL FLOWS

Our discussion of the adjustment mechanism under the gold standard ignored capital flows. In practice, capital flows took place and frustrated adjustment in two ways.

First, countries with a trade deficit sometimes raised domestic interest rates to encourage a capital inflow. A trade deficit no longer implied a payments deficit and monetary outflow. Recession and downward pressure on domestic wages and prices could be avoided.

A trade deficit could persist longer than the idealized account of automatic adjustment suggests.

Second, capital flows explain much of UK economic performance in the nineteenth century. The industrial revolution and market access to a worldwide empire caused a huge UK trade surplus, offset by a huge capital outflow, partly because of heavy investment abroad. However, investment gradually earns interest and profits. Eventually, UK foreign assets were so large that the current account inflow of interest, profits and dividends exceeded the rate at which profitable opportunities for new capital outflows could be found. Recall the basic balance of payments arithmetic:

$$\begin{aligned} \text{trade surplus} + \text{net asset income from abroad} &= \text{current account} \\ &= \text{net outflow on capital account} \end{aligned}$$

The table below shows estimates for the UK for three decades: 1820–30, 1860–70 and 1890–1900. These estimates are in £m in current prices. Because the UK was on the gold standard, prices showed no long-run trend, so nominal values are quite a good guide to real values. However, the UK economy was growing steadily over the period. If we had all the data, it would probably be more informative to show everything as a percentage of GDP. Even so, the table confirms the analysis of the previous paragraph.

Decade average	Current account			Capital	Stock of
	Total	Trade	Debt interest	Outflows	Net foreign assets
1820–30	7.5	3.5	4.4	6.5	110.7
1860–70	36.7	12.1	26.2	31.1	690.7
1890–1900	51.6	−44.1	98.4	46.5	2398.2

Source: Elise Brezis, ‘Foreign capital flows in the century of Britain’s industrial revolution’, *Economic History Review* 48, no. 1 (1995): 46–67.

After decades of current account surpluses, by 1860–70 the UK had stockpiled £690 million of net foreign assets already generating debt interest from overseas of £26 million per annum. In that decade, annual opportunities for new profitable capital outflows were only £31 million. The UK had almost reached the point at which money was flowing in faster than it could successfully be reinvested.

By 1890–1900 this *was* the case. Foreign assets of £2398 million generated an interest inflow of £98 million per annum, much larger

than new capital outflows of £46 million, leaving a net inflow of £52 million. To reconcile the balance of payments, without truly massive changes in gold reserves, it required a trade deficit of £44 million per annum.

In the preceding year, a net inflow of money had raised domestic prices and wages, making UK producers uncompetitive, the market mechanism reconciling current and capital accounts of the balance of payments. External balance meant a big trade deficit plus large net inflows of property income. The monetary adjustment mechanism of the gold standard made inevitable a UK trade deficit in the late nineteenth century. It was not necessarily the result of laziness or decadence, as Victorians believed at the time. An understanding of international monetary economics would have avoided the angst experienced in late Victorian Britain.

The gold standard in action

The gold standard had a big benefit and a big drawback. By tying the money supply to gold, it ruled out persistent large-scale money creation, and hence ruled out persistently high inflation. The UK price level in 1914 was the same as in 1816. In between, in some decades prices rose by 20 per cent and in others they fell by 20 per cent.

However, since monetary policy was dictated by the flow of gold implied by the balance of payments, interest rates could not be set independently for domestic purposes to counter an anticipated boom or slump. Instead, monetary policy had to wait for flows of gold to change the money supply and thus change domestic prices and wages. It could take years to adjust fully to a large fall in aggregate demand. During the gold standard, individual economies had long and deep recessions, as well as sustained booms.

We discuss the Eurozone later in the chapter. Our look at the gold standard already gives three helpful hints. First, permanently fixing nominal exchange rates does not permanently fix real exchange rates. Eventually, competitiveness can change because domestic prices can adjust relative to foreign prices. Second, by curtailing the role of monetary policy, a monetary union raises the significance of fiscal policy for individual member countries wishing to manage aggregate demand independently of

the rest of the monetary union. Third, monetary union is therefore easier when wage flexibility in member states is greater.

26.2 An adjustable peg

An **adjustable peg** is a fixed exchange rate, the value of which may occasionally be changed.

In operation during 1945–73, the most famous example of an **adjustable peg** was called the Bretton Woods system, after the small town where US and UK officials met in 1944 to agree its details. Because countries agreed to use dollars as well as gold as foreign reserves, the system was also called the dollar standard.

Each country fixed its exchange rate against the dollar. The price of gold was fixed in dollars. Currencies were convertible against dollars or gold, which together were foreign exchange reserves. At the fixed exchange rate, central banks were committed to buy or sell domestic currency for foreign exchange reserves. They intervened in the forex market to defend the exchange rate against the dollar.

Unlike the gold standard, the dollar standard did not require 100 per cent forex reserve backing for domestic currency. Governments could print as much money as they wished. The designers of the Bretton Woods system feared that the world gold supply could not increase quickly enough to keep up with the rising demand for money that they hoped would accompany post-war prosperity.

Giving governments the discretion to print money solved that problem but created two others. First, it inhibited the adjustment mechanism built into the gold standard, in which countries with a balance of payments deficit lost gold and, second, it reduced their domestic money supply, bidding down prices and boosting competitiveness. Under the dollar standard, countries with a payments deficit lost money, but the government could print more money. This prevented higher unemployment in the short run, but also prevented long-run adjustment by stopping the fall in prices that raised competitiveness.

If the balance of payments deficit persisted, the country ran out of foreign exchange reserves, and had to devalue its exchange rate to raise

competitiveness, removing the underlying imbalance in international payments.

Speculators faced a one-way bet. When a country was in payment difficulties, its exchange rate would be devalued sooner or later. Speculators might as well bet on devaluation, since the exchange rate was unlikely to appreciate. Sometimes speculative pressure made devaluation happen earlier because of a financial account outflow. Foreseeing this difficulty, the architects of the Bretton Woods system decided to make private capital flows illegal.

Perfect capital mobility implies interest parity. Interest differentials must be offset by expected exchange rate changes to equate expected returns in different currencies. Fixed exchange rates imply expected exchange rate changes are zero. Hence, interest rates have to be equal. Countries cannot retain the sovereignty to set interest rates.

Fixed exchange rates, perfect capital mobility and monetary sovereignty are the impossible triad. All three cannot co-exist.

In 1944 the architects of Bretton Woods decided that fixed exchange rates were important, but that countries were not ready to surrender monetary sovereignty. Hence, capital mobility had to be suspended. Capital flow controls were severe until 1960, when controls on long-term capital flows were relaxed. After the adjustable peg was abandoned in 1973, the need for capital controls diminished. Capital controls were gradually dismantled and integration of global financial markets intensified.

The dollar standard had a second drawback. It led to a world of sustained inflation. Dollars had become the world's medium of exchange. A US payments deficit could be financed by printing more dollars. In the mid-1960s US payments deficits increased, partly because of heavy military spending in Vietnam. The supply of dollars rose rapidly. Raising the world's money supply led to inflation throughout the trading world.

26.3 Floating exchange rates

The **purchasing power parity (PPP)** path of the nominal exchange rate is the path that keeps the real exchange rate constant. Nominal exchange rate changes offset inflation differentials between countries.

Pure floating implies that forex markets are in continuous equilibrium without government intervention to use forex reserves. Reserves are constant. There is no external mechanism to change domestic money supply. The balance of payments is exactly in balance. In the long run, exchange rates adjust to achieve external balance. This determines the real exchange rate that has to prevail in the long run. Domestic and foreign monetary supplies determine the domestic and foreign price levels. Given these, there is only the path of the nominal exchange rate that delivers the real exchange rate required for external balance when domestic economies are also at internal balance.

In long-run equilibrium, nominal exchange rates then obey **purchasing power parity (PPP)**.

When capital mobility is high, neither external balance nor PPP need hold in the short run. Chapters 24 and 25 discussed how exchange rates adjust to achieve interest parity and prevent massive one-way capital flows. In the short run, what matters is not balancing current account flows but the need to balance the potentially much larger financial account flows that might occur when international capital is highly mobile.

In the long run, there is no conflict. Once real interest rates return to their long-run equilibrium level, nominal interest differentials reflect inflation differentials; PPP and interest parity can be satisfied simultaneously.

Hence, a floating exchange rate does not provide continuous short-run insulation against large changes in competitiveness. However, floating exchange rates offer no one-way bet to speculators, since new information can make exchange rates jump up or jump down. Floating helps prevent massive capital flows that cause acute problems for macroeconomic management. Floating exchange rates are also the fallback exchange rate regime when countries cannot agree what other regime to adopt.

A managed float

In a **managed float**, central banks intervene in the forex market to try to smooth out fluctuations and nudge the exchange rate in the desired direction.

Under a free float there is no central bank intervention in the forex market. The forex reserves are constant, the balance of payments is zero, and the net monetary inflow from abroad is also zero. In practice, exchange rates

have rarely floated absolutely freely since 1973, when the Bretton Woods adjustable peg was replaced by a floating exchange rate regime.

Intervention may smooth day-to-day exchange rate fluctuations; in the long run it probably makes little difference to the path the exchange rate follows. Central banks have large stocks of foreign exchange reserves which they could dump on the foreign exchange market to try to alter the equilibrium exchange rate. But nowadays speculators have even larger funds at their disposal.

26.4 Speculative attacks on fixed exchange rates

National policy makers hate to admit that national sovereignty is being eroded. Capital mobility rose sharply after 1980. Under floating exchange rates, potential capital flows affect the exchange rate immediately. Under fixed exchange rates, capital flows could cause the existing peg to be abandoned.

In 1990 the UK pegged its exchange rate against EU currencies in the Exchange Rate Mechanism (ERM), the precursor to the euro. In 1992 the UK was forced to abandon this exchange rate peg against its European neighbours. By floating its exchange rate, the UK regained the ability to set its own interest rates. Lower UK interest rates brought an end to its recession, but led to sharply lower levels of the sterling exchange rate.

More recently, supposedly pegged exchange rates were successfully attacked in Mexico (1994), many Asian countries (1997), Brazil (1999) and Argentina (2002). When the speculators have more money than the central bank, the peg does not always survive.

With high capital mobility, modern crises are not caused directly by trade deficits but by financial account outflows, often reflecting a view that current policy is unsustainable. Raising interest rates to defend the currency may be unconvincing if the domestic economy obviously cannot stand the pain of high interest rates for long.

A **speculative attack** is a large capital outflow. If successful, it causes a devaluation.

There are several interpretations of a **speculative attack**. It may correct a policy mistake. If a country has such a large budget deficit that it needs to finance by creating money, it is bound to have inflation. Promising to peg the nominal exchange rate then makes little sense. A speculative attack merely forces policy makers to switch to a more realistic exchange rate regime; namely, floating.

A second interpretation is that there are two possible equilibrium exchange rates. Without any attack, the original peg survives. The exchange rate is a little overvalued, but the cost of devaluating (likely to raise inflation expectations) outweighs the cost of a little uncompetitiveness. However, once attacked, the cost of repelling the attack must be added to the scales. It may tip the balance, making it optimal now to accept defeat and take the (temporary) advantage of higher competitiveness that the devaluation achieves. Whether the peg survives or not depends entirely on whether speculators decide to attack.

Successful attacks may do a lot of damage. When domestic banks have borrowed in foreign currency, the domestic value of their debts rises when the exchange rate falls. This may bankrupt the banks and cause a widespread loss of confidence. If a country wants to be less vulnerable to attack, what can it do?

Repelling boarders

Capital controls prohibit, restrict or tax the flow of private capital across currencies.

Three responses have been adopted. First, try to reduce capital mobility, making it easier to defend fixed but adjustable exchange rate pegs. This was the solution adopted by those designing the Bretton Woods system after the Second World War. Private capital flows were outlawed by **capital controls**.

Capital controls make it easier to defend pegged exchange rates. However, from the 1970s onwards, controls were progressively dismantled as a global financial system was created. It became harder and harder to enforce controls – smart bankers found offshore ways to do the same business.

A tiny ‘Tobin tax’ on financial transactions may help. Paying a tiny tax on a ten-year investment is trivial; the same tax on holding a foreign asset for

two hours takes away all the profits. A Tobin tax mainly hits short-term ‘hot money’.

Capital controls have been used quite successfully in Chile, and were introduced by Malaysia in 1997 after its currency was attacked. Whether the global economy is consistent with widespread controls is doubtful. Small emerging markets can probably use them. The more highly integrated a country is with the world’s financial markets, the harder it is to use capital controls.

If capital controls are not the answer, the exchange rate regime has to become more robust. Pegged rates are an uncomfortable halfway house: usually pegged, sometimes adjustable. While pegged, the central bank has to defend them, even when a one-way bet is emerging. Because they are not completely pegged, speculators can win in the end.

A **currency board** is a constitutional commitment to peg the exchange rate by giving up monetary independence.

The safer extremes may be float freely or peg completely. Thus, a second solution for repelling boarders and avoiding spectacular exchange rate crashes is simply to float. Let the speculators punch thin air. They can take the currency down, but, if it was for no good reason, the currency will probably come up again.

The alternative is to make the peg much more credible, akin to the old gold standard. A popular device is a currency board.

A **currency board** removes the ability of the central bank to create money. Balance of payments surpluses (deficits) are the only source of expansion (contraction) of the monetary base. Countries using currency boards at some point within the last two decades include Estonia, Bulgaria and Argentina.

Like all commitments, it hurts when it has to take the strain. The country cannot use interest rates for domestic objectives. To live comfortably with a fixed nominal exchange rate, it needs to avoid higher inflation than its trading partners. If it has a fiscally irresponsible government, it will require money creation and inflation that will eventually get it into trouble. However, knowledge of the monetary commitment may induce the fiscal authorities to behave responsibly, as for example in Estonia prior to EU entry.

No solution is ideal. A perfect one would have been adopted everywhere long ago.

MATHS 26.1

A SIMPLE MODEL OF A BALANCE OF PAYMENTS CRISIS

Professor Paul Krugman won the Nobel Prize for his pioneering research on international trade, but his first great paper was on balance of payments crises. Here is a simple version of his argument.

Assume a classical economy with full price flexibility. Output is always at potential output, assumed to be constant. Fiscal policy is out of control, with a constant exogenous budget deficit financed only by printing money. Initially, the country has a fixed exchange rate E . If P is the domestic price, P^* the foreign price level and S the fixed equilibrium real exchange rate, then $EP/P^* = S$. We can use a little calculus: differentiating with respect to time, and assuming that foreign prices are constant, the instantaneous rates of change of domestic variables are:

$$de/dt = -dp/dt = -\pi \quad (1)$$

Thus e and p are the logarithms of E and P , d/dt is the instantaneous rate of change and inflation π is the percentage rate of change of prices. The nominal exchange rate is depreciating at the rate of inflation in order to keep the real exchange rate constant.

Domestic money M is the liability of the central bank, backed by two corresponding assets, its holdings of domestic debt D and foreign reserves R :

$$M = D + R \quad (2)$$

Assume an exogenous rate of growth of D to finance the exogenous budget deficit. With an exchange rate peg, equation (1) implies domestic inflation is zero, in order to maintain the required equilibrium real exchange rate. With prices constant and real output fixed at potential output, real money demand is constant. With no inflation, nominal money M is also constant. Hence, in this initial exchange rate regime, domestic debt D is growing over time to

finance new budget deficits. Equation (2) implies that foreign reserves R are shrinking steadily.

Before Krugman's 1979 paper, economists assumed that, when reserves fell to zero, the country would abandon the peg and float its currency. With D still continuing to grow but R now zero, the money supply would be growing. In this classical model, everything is constant in real terms. Money growth generates inflation and nominal exchange rate depreciation, which was feasible under floating exchange rates but precluded by the previous fixed exchange rate regime and the need to preserve competitiveness.

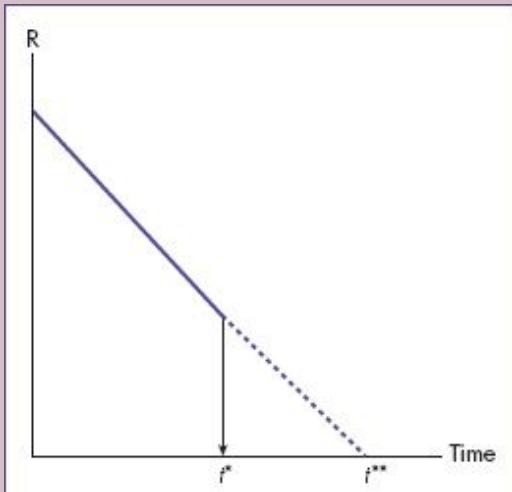
Consider what happens at the instant of the switch from fixed to floating. Inflation jumps up and hence nominal interest rates jump up by this amount – in the classical world, the real interest rate is constant at the level consistent with potential output. Hence real money demand must instantaneously jump downwards to reflect the jump up in the cost of holding money. If reserves R are zero, and D is evolving smoothly to finance the budget deficit, there can be *no* jump in nominal money. Hence, the jump in real money must be achieved by a jump up in the price level, which is certainly flexible enough to do so.

This requires a jump down in the nominal exchange rate, to preserve the equilibrium real exchange rate. However, this jump is predictable in advance since everyone sees the reserves running out and the regime switch coming. Efficient asset markets do not foresee asset price jumps; they should already have acted. This is not the right answer.

If the required jump in M/P cannot be achieved by jumps in P , and hence also in E , then it has to be achieved by a jump in M . Going back to equation (2), in which domestic debt D is growing smoothly to finance the ongoing budget deficit, the required jump in money supply is therefore achieved by a jump in reserves!

At the instant of the regime switch t^* , there is no jump in the price level or the nominal exchange rate. But with higher inflation and nominal interest rates from now on, reserves R jump down to zero in order to facilitate a jump down in real money supply in line with the jump down in real money demand, caused by expecting higher inflation from now on as the economy floats and prints money. There is a successful ‘speculative attack’ on the fixed exchange rate regime that wipes out its remaining reserves in an instant at time t^* before

the time t^{**} that reserves would otherwise have gradually fallen to zero in the fixed exchange rate regime. Once we know how large the rise in inflation will be when the float begins, we know the extent of the change in money supply required. The attack takes place when the country has just enough reserves left to accomplish this jump by having the reserves fall instantly to zero.



Source: P. Krugman, ‘A model of balance of payments crises’, *Journal of Money, Credit and Banking*, 11, 2003 (1979): 311–325.

CASE 26.1

WORLD ECONOMY IN A CUL DE SAC?

After the slump of 2008/09, the world economy partly recovered in 2009/10 after the largest monetary and fiscal expansion in peacetime, not only in the US and Europe but also in China. Fiscal stimulus cannot be sustained at this level. What happens when fiscal policy has to tighten again, as is now happening?

Recall the national income accounting identity that holds as a matter of definition:

$$\text{Injections } I + G + X = S + T + Z = \text{Leakages}$$

Hence,

$$[X - Z] + [G - T] = [S - I]$$

In practice, external deficits are never that large relative to GDP. Hence, any dramatic desire of the private sector to run a substantial surplus ($S > I$) to rebuild its finances inevitably has to be matched by large government deficits. This is necessarily true for the world as a whole, since the world cannot collectively run an external surplus or deficit. Government finances are unlikely to improve dramatically until the private sector has become comfortable again with its debt in relation to its income. Renewed private spending (higher consumption and investment, lower saving) is as much a part of the solution as any direct boost from fiscal policy.

The OECD identified a private sector surplus, in the aftermath of the crash, of around 7 per cent of GDP, both for the Eurozone and for the OECD as a whole. Given previously low levels of saving, this reflected a large rise in the private sector saving rate. Equivalently, it was a large fall in private sector spending, reducing aggregate demand. Without the boost from looser fiscal policy, output and employment would have fallen by even more.

If governments are to tighten fiscal policy, as they must, without causing a further collapse of aggregate demand, private spending must be jump-started again. This cannot be achieved by further monetary loosening. Interest rates are already near zero. In the optimistic scenario, credit growth is somehow restored in the rich countries: both firms and households, anticipating future economic growth, are prepared to start borrowing and spending again. In the pessimistic scenario, the private sector and its potential creditors remain scared of the economic future once fiscal policy cuts back. Private saving remains high, consumption and investment demand remain low, and the paradox of thrift (see Chapter 16) stifles economic recovery.

The rich, but now highly indebted, OECD countries can ‘solve’ their debt/GDP problem either by reducing their debt or by increasing their income. The latter is much less painful if it can be achieved. One possible source of higher aggregate demand is export demand from emerging markets such as China; since emerging markets have recovered much more quickly, they have not accumulated nearly so much government debt during the crash. This is why OECD politicians are so concerned about the undervaluation of the exchange rates of emerging market economies – and corresponding overvaluation of exchange rates of OECD countries. This reduces the

scope for the US, Europe and Japan to enjoy substantial export booms to emerging markets as these latter economies resume rapid growth.

If the mature economies do not experience this export-led growth, the next-best bet is probably a resurgence of private and public investment. Although this would require further borrowing today, it would not only add to aggregate demand in the short run but also boost aggregate supply in the longer run as this new capital stock became available for production.

Professor Larry Summers, former head of Harvard University and Assistant Treasury Secretary in the Clinton Administration, has argued that, since nominal interest rates are very low, real interest rates will be negative for some time. In such circumstances, the cost of borrowing is effectively negative. If used not for consumption but for productive investment that generates future tax revenue, it actually enhances government solvency not reduces it.

In January 2013, the IMF's update to its *World Economic Outlook* projected a gradual upturn in world growth in 2013, as shown in the table below, though at a slower rate than previously forecast.

Growth rates of world output (% pa)	2012	2013	2014
World	3.2	3.5	4.1
BRICS			
Brazil	1	3.5	4.5
Russia	3.6	3.7	3.8
India	4.5	5.9	6.4
China	6.6	7.1	7.5
US	1.3	1.4	2.2
Euro Area	-0.4	-0.2	1.0
UK	-0.2	1.0	1.9

Sources: adapted from M. Wolf, 'The world economy has no easy way out of the mire', *Financial Times*, 23 February. © The Financial Times Limited 2010; IMF, *World Economic Outlook, Update*, January 2013.

Where the government is weak at home, invoking external pressures may be a useful tactic. But a tough government may be able to institute

domestic forms of commitment (appointing a tough governor of the central bank or giving that bank greater independence from government control) that still leave the choice of exchange rate regime determined by other factors.

26.5

Antecedents of the Eurozone

In the **Exchange Rate Mechanism (ERM)**, each country fixed a nominal exchange rate against each other ERM participant. Collectively, the group floated against the rest of the world.

Because neighbouring small open economies tend to trade a lot with one another, they dislike volatility in the exchange rates between them. Peugeot got upset if Fiat suddenly gained a 20 per cent price advantage simply because the Italian lira had depreciated against the French franc. In 1979 the members of the European Community set up the European Monetary System (EMS), a system of monetary and exchange rate co-operation in Western Europe. Its most important aspect was the **Exchange Rate Mechanism (ERM)**.

Each country in the ERM could let its exchange rate fluctuate within a band of $62\frac{1}{4}$ per cent of the parities it had agreed to defend.¹ When the currency hit the edge of a band, all central banks in the ERM countries were supposed to intervene to try to defend the parity. Realignments of the fixed exchange rates against partner countries were possible but had to be unanimously agreed upon by participants of the ERM.

Initially, the target exchange rates had to be realigned about twice a year – higher inflation countries needed to devalue frequently in order to restore competitiveness.

Since high-inflation countries needed regular devaluations, which required the consent of other ERM members, by withholding such consent it was possible for low inflation countries such as Germany to put pressure on the high-inflation countries. Monetary policies steadily converged, especially after 1983, when realignments were much less frequent. Between 1987 and 1991 there was no realignment at all.

The role of Germany, as the largest country, was crucial. Determined never to repeat the hyperinflation of 1923, Germany could be relied on to keep inflation low. Low inflation expectations gradually emerged in Spain

and Italy not because their own policies were suddenly more credible but because Germany might block the devaluations that would be needed if inflationary policies ever re-emerged.

Why the ERM survived

How did the ERM survive in the early years before inflation converge was achieved? First, with an exchange rate band $2\frac{1}{4}$ per cent either side of the central parity, high inflation countries had some room to depreciate within the band. When they got near the bottom of the band, devaluation soon followed. The early ERM was largely cosmetic.

Second, most countries initially had capital controls preventing big financial account flows. However, to create a single market in the EU, in 1987 countries committed themselves to the steady removal of capital controls. ERM members had a common interest rate, effectively set by Germany.

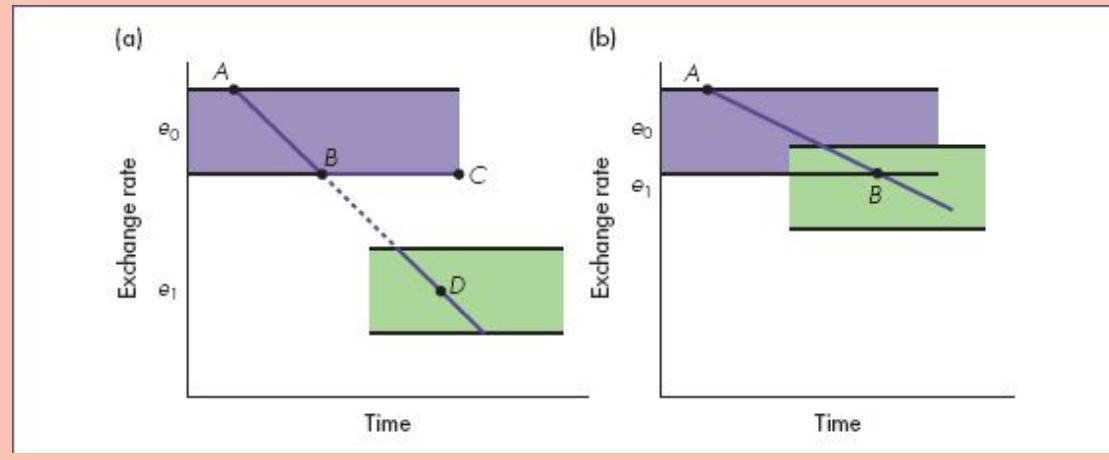
CASE 26.2

CAPITAL CONTROLS AND THE EARLY ERM

For a high-inflation country, figure (a) shows the path *ABD* of nominal exchange rate depreciation to maintain real competitiveness. Initially, the exchange rate is pegged within a band around central parity e_0 . When the exchange rate hits the bottom of the band at *B*, central banks intervene to try to defend the band. As time elapses, the exchange rate moves along *BC*. With continuing inflation, competitiveness is now being eroded. Eventually this prompts a devaluation of the central parity from e_0 to e_1 , so the whole band shifts down. The actual exchange rate jumps from *C* to *D* on the day of the realignment.

This is a one-way bet. As the exchange rate moves along *BC* nobody is expecting a sharp appreciation! Only capital controls prevent a massive outflow on the financial account to avoid the imminent capital loss on holding the currency. If capital controls had been removed in the early 1980s, there would have been an immediate crisis. Figure (b) explains why this did not occur after 1986.

By the mid-1980s, inflation convergence within the ERM meant even Italian inflation was nearly down to German levels. Slow depreciation of the lira could offset extra Italian inflation. The line is flatter in figure (b) than in (a). In figure (b) when the parity is devalued from e_0 to e_1 , the actual exchange rate at B is inside both the old band and the new band. No jump is required in the exchange rate, and there is no one-way bet for speculators. A small interest differential will compensate them for the gradual depreciation.



From ERM to EMU

Once capital controls were removed, sooner or later speculators would attack the pegged exchange rates of the ERM. One solution was to go forward rapidly to completely fixed exchange rates. A monetary union need not have a single currency. English and Scottish currencies circulate side by side in Edinburgh. What matters is that the exchange rate is certain and that a single monetary authority sets the interest rate for both. A monetary union has permanently fixed exchange rates within the union, an integrated financial market and a single interest rate.

The [Maastricht criteria](#) for joining EMU said that a country must already have achieved low inflation and sound fiscal policy.

The Treaty of Maastricht in 1991 set out the criteria for Eurozone membership. Any remaining capital controls were to go, and prospective members had to satisfy the [Maastricht criteria](#). The Eurozone was launched in 1999, with most EU members satisfying these criteria. Greece initially did not meet the criteria, but joined later; Denmark and Sweden

opted out. Monetary policy in the EMU was to be set by an independent European central bank, whose principal goal was to be price stability.

The fiscal criteria said budget deficits must not exceed 3 per cent of GDP and that the debt/GDP ratio should not be over 60 per cent. Tight fiscal policy should mean there was little pressure on the central bank to print money to bail out fiscal authorities. The financial crisis of 2008 led to widespread increases in budget deficits as governments bailed out their failing banks. When it came to the crunch, the Maastricht criteria were ignored.

The Maastricht deal reflected the balance of power in the negotiations. At the time, Germany ran the EMS and trusted itself to do so in its own interests. Germany would give up such a good position only if the EMU was going to be super-safe. The Maastricht criteria were the price of getting Germany on board.

Sterling and UK membership

Why was the UK reluctant to join both the ERM and the Eurozone? First, until the late 1980s, North Sea oil made sterling behave differently from other European currencies. As UK oil production slowed down, this objection evaporated.

Second, whereas the core countries of Europe are now very integrated with one another, offshore UK, like other peripheral EU states, is less integrated with the rest of Europe. A common policy is less suitable. However, Table 26.2 confirms that the EU has become the UK's principal trade partner. If this trend continues, the UK may eventually find it advantageous to join.

Table 26.2 UK trade patterns (% of UK trade)

	EU	North America	Rest of world
1970	34	17	49
2005	62	11	27

Sources: UN, International Trade Statistics; www.statistics.gov.uk.

The growth of emerging market economies is forcing all developed economies to trade more with emerging markets. If China and India boom sufficiently, UK trade with Europe may actually fall a bit. However, there is considerable evidence that countries trade considerably more with their

nearest neighbours. Euro area trade is always likely to remain the single largest focus of UK trade.

Third, the UK has greater macroeconomic sovereignty: it seems to have more to lose. ERM countries had already allowed the ‘single’ interest rate to be set by Germany alone; sterling floated during the entire period except for the two years of its ERM membership during 1990–92. After the financial crash, the UK happily allowed sterling to depreciate to cushion the recession. The Bank of England resorted to quantitative easing without having to consult EU partners, and bond markets have been content with low UK interest rates, knowing that UK fiscal solvency is ultimately underpinned by the ability to create money if necessary.

Finally, Black Wednesday (16 September 1992) made it hard for UK politicians to enthuse about the euro. The UK joined the ERM in 1990 to combat rising inflation at the end of the Lawson boom. Unfortunately, this coincided with German reunification. Big subsidies to East Germany caused German overheating, forcing the Bundesbank to set high interest rates to cool the German economy. Interest rates high enough to do this job were far too high for Germany’s partners in the ERM. This provoked the crisis of 1992–93. The UK left the ERM, slashed interest rates and depreciated sharply, allowing rapid output recovery.

26.6 The economics of the euro

An **optimal currency area** is a group of countries better off with a common currency than keeping separate national currencies.

Mundell, and the economists who came after him, identified three attributes that might make countries suitable for a common currency. First, countries that trade a lot with each other may have little ability to affect their equilibrium real exchange rate against their partners in the long run, but face temptations to devalue to gain a temporary advantage. A fixed exchange rate rules out such behaviour and allows gains from trade to be enjoyed.

Second, the more similar the economic and industrial structure of potential partners, the more they will face common shocks that can be dealt with by a common monetary policy. It is country-specific shocks that pose difficulties for a single monetary policy.

Third, the more flexible are labour markets within the **currency area**, the more easily any necessary changes in competitiveness and real exchange rates can be accomplished by (different) changes in the price level in (different) member countries.

Conversely, countries gain most by keeping their monetary sovereignty when they are not that integrated with potential partners, have a different structure and hence are likely to face different shocks, and cannot rely on domestic wage and price flexibility as a substitute for exchange rate changes.

To these purely economic arguments, we should add an important political argument. Currency areas are more likely to work when countries within the area are prepared to make at least some fiscal transfers to partner countries. In practice, this cultural and political identity may be at least as important as any narrow economic criteria for success.

CONCEPT 26.2

SOVEREIGNTY AND MONETARY UNION

Perfect capital mobility undermines **monetary sovereignty**. If interest rates are set to maintain the pegged exchange rate, they cannot be set independently to meet the needs of the domestic economy.

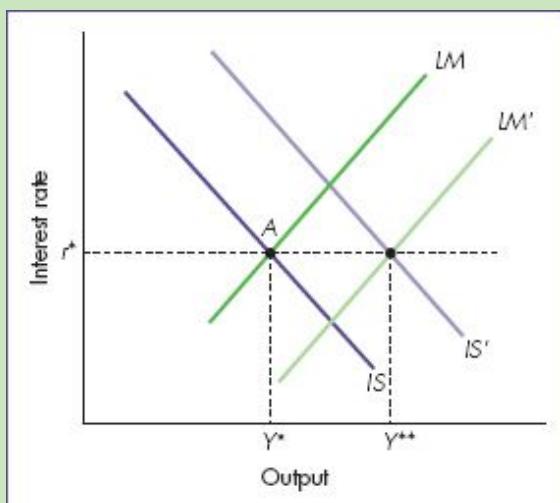
In 1999 Chicago professor Robert Mundell won the Nobel Prize for helping invent open economy macroeconomics. He was the first to realize that openness in output and factor markets creates powerful pressure for monetary union. He also showed what it would be like for a small country to try to hang on to **monetary sovereignty** when international capital mobility is high.

The figure below shows a pegged exchange rate. The *IS* schedule shows the usual relationship between interest rates and output consistent with goods market equilibrium. A small country can peg its exchange rate only by matching the foreign interest rate r^* . We show this as a horizontal line. The money supply adjusts to make sure this is always the domestic interest rate. Initial equilibrium is at *A*.

As in Figure 25.2, any attempt to change the money supply, and hence interest rates, causes an immediate capital inflow or outflow on the financial account until the money supply and interest rates are restored to r^* . For a small open economy with a pegged exchange rate, monetary policy is powerless.

A fiscal expansion shifts IS to $IS\epsilon$. There is a big short-run effect on output, from Y to $Y\epsilon$, since interest rates cannot rise to dampen the expansion. Monetary policy is forced to create additional money supply to accommodate the extra money demand when output rises. We can think of the horizontal line for interest rates as being achieved by a shift in the implicit LM schedule from LM to $LM\epsilon$. In fact, we may as well regard the horizontal line as the LM schedule itself. The entire role of monetary policy is to peg the interest rate at r^* .

Y^* is potential output. If a demand shock shifts IS to $IS\epsilon$, potential output is not restored by induced changes in interest rates as in a closed economy. Interest rates remain at r^* . Rather, higher prices reduce competitiveness and hence shift $IS\epsilon$ leftwards again. Under a pegged exchange rate, induced changes in IS schedules restore output to full capacity. If this takes too long, fiscal policy must shift $IS\epsilon$ back to IS .



Is Europe an optimal currency area?

Those who have studied the structure of national economies, and the correlation of shocks across countries, generally reach the following

conclusions. First, Europe is quite, but not very, integrated. Second, there is a clear inner core of countries – France, Germany, Netherlands, Belgium, Luxembourg and Austria – more closely integrated than the rest.

However, the act of joining the Eurozone changes the degree of integration, possibly quite substantially. A common currency, by eliminating a source of segmentation into national markets, increases integration. Moreover, there is evidence that countries that trade a lot have more correlated business cycles. And countries which belong to currency unions tend historically to trade much more with each other than can be explained simply by the fact that their exchange rates are fixed.

These bits of evidence imply that it may be possible to start a currency union before the microeconomic preconditions are fully in place. The act of starting speeds up the process. Even so, a decade of the Eurozone did *not* prepare all its members for the severe strain that would be imparted by the financial crash and its fiscal aftermath.

The Stability Pact

The Stability Pact, ratified by the Treaty of Amsterdam in 1997, confirmed that the Maastricht fiscal criteria would not merely be entry conditions for EMU but would also continue to apply after countries joined the monetary union. Some EMU members had debt/GDP ratios of close to 100 per cent. Reducing these towards 60 per cent might take decades. The focus was on the 3 per cent ceiling for budget deficits.

In principle, countries exceeding the limit had to pay fines unless their economy was in evident recession. Thus countries had to wait for output to fall before they were allowed to expand fiscal policy by having deficits above the ceiling of 3 per cent of GDP.

The pact did not preclude countries from using fiscal policies more vigorously, as they most certainly did in 2008/09, when they were clearly in recession. The corollary is that, ideally, they should aim for something more like budget balance in normal times and for budget surplus during periods of boom. Then they still have room to increase deficits in times of trouble without exceeding the 3 per cent ceiling.

Note that if budgets are roughly in balance over the business cycle, but output grows for ever, debt/GDP ratios should exhibit trend decline, whatever their cyclical behaviour. This may eventually lead to the tough conditions of the Stability Pact being eased.

26.7

The European Central Bank

The single monetary policy is now set in Frankfurt by the European Central Bank (ECB). National central banks have not been abolished, but the board of the ECB sets the interest rate on the euro.

The ECB mandate says its first duty is to ensure price stability, but it can take other aims into account provided price stability is not in doubt. In press conferences, ECB officials have emphasized that their interest rate decisions should be interpreted largely as the pursuit of price stability. Neither the financial markets nor academic economists are entirely convinced. ECB behaviour looks as if they pay some attention to output gaps as well as inflation. Empirically, a Taylor rule explains their behaviour quite well: interest rates are changes to reflect deviations of both inflation and output from their underlying target levels.

Empirically, it is always difficult to distinguish between a central bank that actually cares about output as well as inflation (however much it says its only concern is inflation) and a central bank that cares about current and future inflation, and uses output gaps as one means of forecasting future inflation. For whichever reason, a Taylor rule works well in capturing empirically how central banks actually behave in setting interest rates.

Despite the fact that its behaviour is largely explained empirically by a Taylor rule, the ECB describes its monetary strategy as pursuing two intermediate targets – the ‘twin pillars’. The first pillar is a monetary target, the growth rate of the M3 measure of nominal money. The second pillar is expected inflation. The ECB insists that it takes both pillars into account in setting interest rates in the Eurozone.

Figure 26.1 shows the interest rate decisions of the ECB, the evolution of inflation and the rate of nominal money growth. It is easy to see how the rise and fall of actual and expected inflation led to the rise and fall of interest rates in the Eurozone. It is very hard to detect any clear correlation between nominal money growth and interest rate decisions.

It is precisely the volatility in both money demand and the money supply behaviour of banks that has made most central banks abandon monetary targeting in favour of flexible inflation targets or Taylor rules. The ECB continues to insist that monetary targets have an important role because it wants to emphasize continuity with the Bundesbank, which uses monetary targets.

Countries such as the UK and the US engaged in substantial quantitative easing, sharply increasing the supply of narrow money in order to prevent a collapse of the banking multiplier inducing a corresponding fall in broad money. Figure 26.1 shows that the ECB never allowed broad money to fall. On the other hand, its growth rate since 2009 has been small. Traditional German fears about the consequences of printing money inhibited the ECB response to the crisis, at least until 2012, when the new governor, Mario Draghi, led a sharp change in ECB policy.

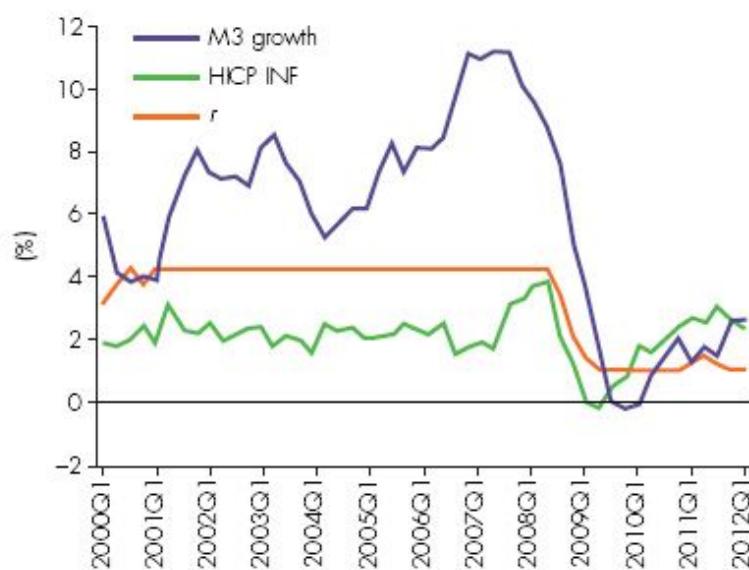


Figure 26.1 The ECB in action, 1999–2012

Note: HICP measures the inflation in harmonized index of consumer prices; annual % growth of board money M3; r is the short-term interest rate controlled by the ECB.

CASE 26.3

THE ECONOMIC CONSEQUENCES OF MR DRAGHI

European policy making has always been a dialogue between the German tradition, informed heavily by the need not to repeat the hyperinflation of 1922/23, and the activist tradition – often including

France, the Mediterranean countries and sometimes the UK – which also stressed the need to prioritize economic growth and low unemployment.

The initial design of the ECB was heavily influenced by German success. It stressed price stability and monetary responsibility. Its first president, Wim Duisenberg (1999–2003), was a Dutch politician and central banker, whose leadership of the Dutch central bank had displayed caution and responsibility. He earned the nickname ‘Mr 15 Minutes’ because, every time Germany changed interest rates, the Netherlands matched the interest rate change immediately, credibly anchoring the Dutch guilder to the German mark. Duisenberg was comfortable with the Germanic approach.

His successor, Jean-Claude Trichet (2003–11), Duisenberg; Trichet; Draghi was an eminent French civil servant, having run the French Treasury in a tough and orthodox manner. The Duisenberg–Trichet years of the ECB were devoted in part to proving that the ECB could be just as trusted with price stability as the Bundesbank had been.

Duisenberg;



Trichet;



Draghi



©INTERFOTO/Alamy; © BRU News/Alamy; © Peter Cavanagh/Alamy

In 2011, Governor Trichet was succeeded by Mario Draghi, who had long experience of European financial diplomacy, most recently as governor of the Bank of Italy. In the 1970s, Draghi had been one of a gifted generation of economics PhD students at the Massachusetts Institute of Economics (MIT) – his peers included Ben Bernanke (now governor of the US Federal Reserve), Olivier Blanchard (now chief economist of the IMF), Paul Krugman (Nobel Laureate in economics) and Ken Rogoff (Harvard professor, former chief economist of the IMF). The MIT economics tradition has always emphasized the importance of the real economy as well as monetary orthodoxy.

ECB watchers were unsurprised when Draghi accomplished a change in ECB policy shortly after taking office. In 2011, despite economic stagnation, the ECB under Trichet had twice raised interest rates to mitigate the effects of imported inflation from higher world commodity prices, in contrast to the US Fed and the Bank of England, which had cut interest rates almost to zero. Under Draghi, the ECB quickly followed suit and reduced interest rates again.

In December 2011, Draghi oversaw a €500 billion loan to troubled banks within the Eurozone. Such measures were controversial, since the ECB is not allowed to put European taxpayers' money at risk. Draghi judged that such measures were likely to help economic recovery in Europe, assist the banking system to begin lending again and weaken self-fulfilling prophecies that banks were doomed.

By summer 2012, a new crisis had emerged. Concerns about banking solvency had been transferred to concerns about government solvency. Many Mediterranean governments were paying very high interest rates to borrow. High interest rates, caused by fears of default, then become a cause of fiscal insolvency that makes default, and the break up of the Eurozone, more likely.

In July 2012, Draghi promised that the ECB would do ‘whatever it takes to preserve the euro’, adding ‘Believe me. It will be enough.’

Immediately, bond markets and equity markets began a long rally. Share prices rose, and long-term interest rates came down for the governments under pressure. In December 2012, Draghi was widely proclaimed as man of the year.

It is helpful to understand two constraints that Draghi had to overcome in order to succeed. First, he had to find a vehicle by which the ECB could lend to governments without having to seek explicit approval from European parliaments. The constitution of the ECB prevents it printing money directly to finance government deficits – one of the many checks designed to prevent another hyperinflation.

Draghi's solution was to argue that he could buy bonds in the secondary market (that is, after they have already been issued by governments), not for the purpose of helping government finances but in order to repair the broken transmission mechanism of monetary policy. Unless there is an unbroken chain from money to short debt to long debt to bank lending to firms and households, monetary policy cannot work properly. The credit blockage was used as an alibi for allowing the ECB to buy government bonds.

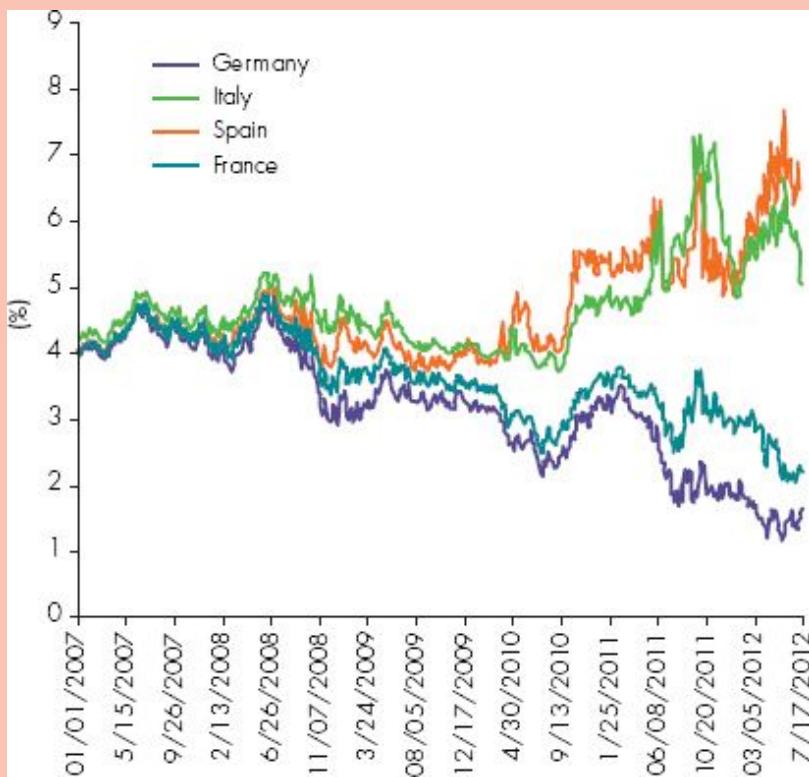
Second, he had to persuade leading politicians, notably German Chancellor Angela Merkel, that his initiative would cost less than the alternative of allowing the Eurozone to collapse. Since monetary union was supposed to be for ever, there were no provisions for how debts would be allocated in the event of a break up. The more people contemplated the possibility, the more complex and costly it appeared to be.

Draghi's initiatives may not solve the crisis permanently, but he bought breathing space for policy makers to undertake more fundamental reforms. Financial markets around the world, less fearful of Eurozone disaster, boomed despite the fact that the US, Europe and Japan were all experiencing feeble growth. The figure below confirms the effect on the interest rates that governments had to pay for long-term borrowing.

France and Germany were regarded as solvent by financial markets. With interest rates low, long-term rates fell in these countries. However, by late 2011 Italy and Spain faced unsustainable interest rates – over 7 per cent a year. Even France faced a bit of a spike. The combination of lower short-term interest rates – the Draghi cuts – and the loan programme to banks caused a substantial fall in long-

term interest rates for countries in fiscal peril. However, the figure also shows that, by July 2012, rates climbed as markets became pessimistic that the Eurozone could survive unscathed. The effect of the Draghi promise to ‘do what it takes’ is shown at the end of the graph – another sharp fall in borrowing rates in Italy and Spain. Similar effects were experienced in Portugal and Greece.

This case study conveys the subtleties of how a monetary union works. There are economic issues caused by the availability of only one monetary policy for the whole diverse region. There are fiscal issues, particularly concerns about the amount of redistribution within the union. In most nation states, the rich are to some extent prepared to pay for the poor. The Eurozone was founded as an attempt to get the benefits of a large single market and common currency *without* any great commitment by the rich northern countries to pay for the poorer countries in the south. Many of the rules were designed to prevent the south ever needing elaborate assistance from the north. These rules proved inadequate because the scale of the financial crash and its consequences were not anticipated. Thereafter, there has been an ongoing tension about the extent to which there will be not just emergency assistance but ongoing transfers from north to south.



Source: IMF, World Economic Outlook, 2012.

Fiscal federalism?

A **federal fiscal system** has a central government setting taxes and expenditure rules that apply in its constituent states or countries.

One reason for the survival of the monetary union that we call the US is its **federal fiscal** structure. When a particular state has a slump, it pays less income tax revenue to Washington and gets more social security money from Washington, without any decisions having to be taken. Automatic stabilizers are at work, courtesy of federal tax rates and federal rates of social security payments. Conversely, a booming state pays more tax revenue to Washington and gets less social security money back.

When a state's income rises by \$1, the state pays an extra 30 cents in income tax and gets 10 cents less in social security. Conversely, when state income falls by \$1, the state pays 30 cents less in federal taxes and gets an extra 10 cents in social security. Originally, economists thought that this meant each state was effectively insured by up to about 40 cents in the dollar. The Eurozone has no federal fiscal structure on anything like this scale. The pessimists concluded that the euro would come under pressure from country-specific shocks.

The idea was correct but the sums were wrong. The original US calculations are relevant to a world in which state incomes are uncorrelated with each other. In practice, the correlation is quite high. Hence, when one state slumps and gets help from Washington, many other states are slumping and also getting help. But this increases US government debt and means *every* state has to pay higher future taxes.

But an individual state could have done that on its own, without membership of the federal 'mutual insurance' club. It could have borrowed in the slump to boost its own fiscal spending, and paid it back later when times were better. Making allowance for this, US states are probably insured by nearer to 10 cents in the dollar than 40 cents.

However, the Stability Pact may have *prevented* individual EMU countries behaving in this way, by restricting their ability to borrow in bad times. In fact, over time the Stability Pact was interpreted more flexibly, having

greater regard for the effect of temporary cycles in temporarily reducing tax revenue.

Since 2010, the Eurozone has thus faced a number of fiscal challenges. First, to what extent could the reputation for fiscal prudence be restored? Second, how could acute differences in the outcomes of different member states be resolved without having a common fiscal policy? Third, could the possible bankruptcy of individual member states be prevented? As in all questions of redistribution, the fortunate have to be willing to pay for the unfortunate. During 2010 German voters made it pretty clear they were not excited about paying for Greece or Portugal, let alone Spain and Italy. The less fiscal support such countries obtain, the more they may be driven to restore their monetary autonomy by leaving the Eurozone altogether.

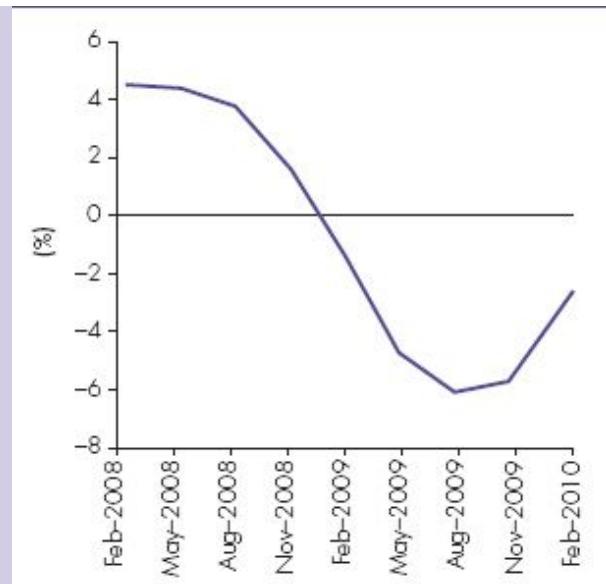
ACTIVITY 26.1

IRISH DEFLATION AND ITS CONSEQUENCES

‘Irish unions agree to link pay rises to efficiency’ (*Financial Times*, 30 March 2010) following a deal between government and public sector trade unions. In exchange for avoidance of compulsory redundancies, unions agreed to flexible work practices and possible pay cuts. The direct effect was to tighten fiscal policy by shrinking public spending. During 2008–10 Irish fiscal policy was tightened by an amount equivalent to 6 per cent of GDP. Ireland aimed to cut its budget deficit from 12 per cent in 2008 back to 3 per cent of GDP by 2014.

The second effect was to reduce the price level in Ireland. If nominal wages are cut, and prices then fall, Irish competitiveness will rise even within the Eurozone. Domestic wage and price reduction is a substitute for nominal exchange rate depreciation.

In the December 2009 budget, public sector workers took pay cuts from 5 per cent for people earning below €30 000 to 10 per cent for those on higher incomes. This was on top of a pension levy in February 2009 that reduced take-home pay by 5 per cent.



Ireland, annual CPI inflation

Source: Central Statistical Office, Ireland.

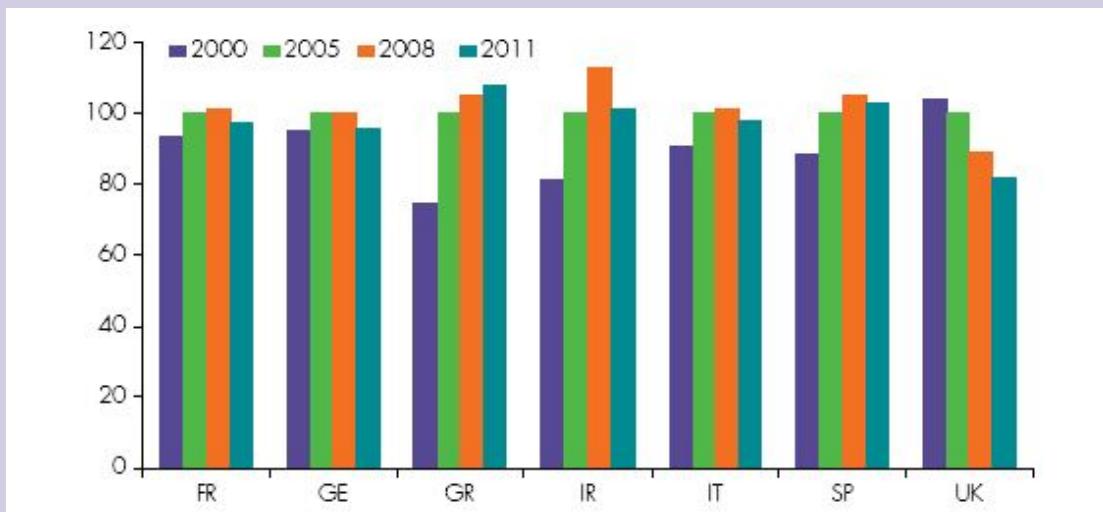
Suppose Irish prices and wages each fall by 10 per cent. How much poorer have Irish people become? The answer depends on the openness of the Irish economy. If imports were only 10 per cent of the size of GDP, real wages would be reduced only a little since most goods are produced domestically and their prices have fallen by the same percentage as nominal wages. Conversely, if imports are 90 per cent of the size of GDP, the nominal wage cut is matched by a price cut on only the 10 per cent of goods produced and consumed in Ireland. Irish residents are worse off by around 9 per cent in real terms. For the same reason, Irish competitiveness has increased substantially.

Because Ireland is a small, open economy, the answer in practice is closer to the latter example than the former. Annual consumer price deflation bottomed out in Ireland at -6.4 per cent in October 2009, but was still -4 per cent as late as January 2010.

The figure below shows what happened to real exchange rates, measured by relative consumer prices in a common currency, treating the index as 100 in 2005. France and Germany had stable competitiveness over time. After 2000, Greece experienced a huge loss of competitiveness. Its entry to the Eurozone saw many prices marked up to German levels, even though Greek productivity levels were much lower. Italy and Spain faced smaller losses in competitiveness. Outside the Eurozone, the UK currency depreciated

sharply downwards after 2008, causing a 10 per cent rise in UK competitiveness.

The real point of the figure is Ireland. Until 2008, Ireland had behaved more like Greece than like Germany. During 2000–2008, the Irish real exchange rate appreciated sharply because of relatively high domestic inflation. During 2008–11, Ireland achieved the *largest* real exchange rate depreciation of any of the countries shown in the figure. From inside the Eurozone, it gained more competitiveness through domestic deflation than was achieved by the UK with its nominal exchange rate depreciation.



Real exchange rates (relative consumer prices in the same currency, 2005 = 100), 2000–2011

Ireland is a small, cohesive, well-educated country, with a relatively flexible labour market. What it accomplished was not easy politically. It shows that there are feasible adjustment mechanisms for a single economy within the Eurozone. The problem in Spain, Portugal, Italy and Greece has been political unwillingness to implement the textbook remedy.

Questions

- (a) Opponents of monetary union argue that it prevents adjustments in competitiveness when required. In each of the following cases, label this problem as important or less important:
- an economy with flexible labour markets
 - an economy that trades only a little with the rest of the world
 - an economy with powerful trade unions

- (iv) an economy that is extremely open to trade with its near neighbours
- (v) an economy with similar industries to its major trading partners
- (vi) an economy whose government always pursues fiscal prudence
- (b) Suppose everyone in the economy simultaneously took a 20 per cent reduction in income, and firms passed on lower costs as lower prices:
- (i) Would people's living standards have fallen?
 - (ii) Does this depend on how open the economy is?
 - (iii) Would it always be better to have the option to depreciate the exchange rate?
- (c) Given that Eurozone countries trade with the rest of the world as well as among themselves, to what extent could a depreciation of the euro resolve the competitiveness problems of Greece, Spain and Portugal?

To check your answers to these questions, go to page 681.

Summary

- Under the **gold standard**, each country fixed the **par value of its currency** against gold, maintained the **convertibility** of its currency into gold at this price and linked the domestic money supply to gold stocks at the central bank. It was a fixed exchange rate regime.
- Without capital flows, countries with a trade deficit faced a payments deficit, lower gold stocks and a lower money supply. Domestic recession then bid down wages and prices, raising competitiveness – an automatic adjustment mechanism. Trade surplus countries faced a monetary inflow, higher prices and lower competitiveness. In practice, this adjustment mechanism was hampered by capital flows.
- The post-war **Bretton Woods system** was an **adjustable peg** in which fixed exchange rates were sometimes adjusted. It was a dollar standard. But domestic money supplies were no longer linked to forex reserves, so the adjustment mechanism of the gold standard was weakened.

- **Purchasing power parity (PPP)** is the path of the nominal exchange rate that would maintain constant competitiveness by offsetting differential inflation across countries. In the long run, floating exchange rates return to the PPP path if no real shocks occur.
- In the short run, the level of **floating exchange rates** is determined largely by speculation. Exchange rates adjust to ensure interest differentials are offset by expected exchange rate changes. This chokes off large speculative flows. In the short run, exchange rates can depart significantly from their long-run level.
- Unlike fixed exchange rates, floating exchange rates can cope with permanent differences in national inflation rates. High-inflation countries have a depreciating exchange rate in the long run. In practice, floating exchange rates also coped with the severe real shocks of the 1970s. Floating exchange rate regimes are more robust than fixed exchange rate regimes.
- Critics of **floating exchange rates** claim they are **volatile** in the short run, which discourages international trade and investment. However, they are volatile because the world is uncertain. Under fixed exchange rates the uncertainty would show up somewhere else, possibly in volatile domestic interest rates to maintain the fixed exchange rate.
- Fixed exchange rates impose financial discipline by preventing a country having permanently higher inflation than the rest of the world. However, there are other ways to commit to low inflation. And fixed exchange rates do not always survive!
- **International policy co-ordination** is hard to implement but allows policy makers to take account of the externalities they impose on each other. It may allow individual governments to commit themselves to policies that would otherwise not be credible.
- The UK was always a member of the **European Monetary System** but belonged to its key feature, the **Exchange Rate Mechanism**, only during 1990–92. The early survival of the ERM arose only partly from

greater co-ordination of monetary policy by ERM participants. Foreign exchange controls and exchange rate bands were also important. After 1983 devaluations became harder to obtain and monetary policies had to converge on the low inflation rate in Germany.

- In abolishing capital controls before 1992, the ERM had already harmonized monetary policy, under German leadership. The UK became an ERM member in 1990, but left in 1992.
- **monetary union** means permanently fixed exchange rates, free capital movements and a single interest rate.
- The **Maastricht criteria** say that EMU entrants, including future ones, must have shown low inflation, low interest rates and stable nominal exchange rates before entry; and must have budget deficits and government debt under control.
- EMU members must continue to obey the **Stability Pact**, which fines countries for excessive budget deficits, except if they are in recession, as in 2009/10.
- In the EMU, a country's competitiveness can change through the slow process of domestic wage and price adjustment. Without a **federal fiscal system**, individual member states may want to keep control of fiscal policy to deal with crises.

Review questions



EASY

- 1 How effective is the managed float regime in the long run?
- 2 What are the advantages and disadvantages of a currency board?
- 3 **Common fallacies** Why are these statements wrong? (a) Floating exchange rates make sure that exports and imports always balance. (b) Fixed exchange rate regimes prevent necessary changes in competitiveness.
- 4 If floating is so great, why did most EU countries join a monetary union in 1999?
- 5 Which of the following statements are correct? Some countries in the Eurozone have suffered speculative attack in 2010 because (a) interest

rates have been unnecessarily high; (b) they have been unable to devalue to boost growth and tax revenues; (c) they can put pressure on richer Eurozone countries to bail them out; (d) they cannot use inflation as a weapon of last resort for deflating away government debt.

MEDIUM

- 6 During the First World War the gold standard was suspended. To pay for the war, Britain printed money and sold off its foreign assets. What do you think happened in 1925 when Britain tried to rejoin the gold standard at the old nominal exchange rate?
- 7 Contrast the dollar standard and the gold standard in terms of the automatic adjustment mechanism.
- 8 You are Finance Minister of Cuba in the new government that has decided to abandon five decades of communism and move as rapidly as possible to a free market. (a) Should Cuban citizens expect rising or falling incomes in the long run? (b) If they implement lifecycle or permanent income approaches to consumption, will they wish to borrow or to lend during this transition? (c) What do you expect to happen to Cuba's international trade balance?
- 9 'What has not flowed in cannot then flow out.' Which exchange rate regime would best insulate a country from capital inflows? What would be the cost of this policy?
- 10 Again, you are Finance Minister of Cuba in the new government that has decided to abandon five decades of communism and move as rapidly as possible to a free market. (a) From now on, will Cuba offer unusually good or unusually bad investment opportunities? (b) What will happen if Cuba pursues a fixed exchange rate policy during this period? (c) What will happen if Cuba follows a floating exchange rate policy? (d) Which of the two would you recommend?
- 11 What are the Maastricht criteria to be met by the Eurozone member countries? Why were these criteria ignored during the financial crisis of 2008?

HARD

- 12 **Essay question** 'Small open economies need fixed exchange rates; large economies need floating exchange rates.' Is this broad generalization correct? Explain why or why not.
- 13 In 2010, Ireland implemented a deliberate policy of linking pay rises to efficiency to induce deflation. (a) How did Ireland plan to achieve

deflation by implementing such a policy? (b) Use an *IS–LM* diagram to show the consequences of negative inflation in Ireland.

|4 Use an *IS–LM* diagram to show what happened to the UK in 1992 after it left the Exchange Rate Mechanism.

- 1 Italy, an especially high-inflation country in 1979, was allowed a band of 6 per cent. By the mid-1980s, it was a matter of honour for Italy not to use this wider band. Spain and the UK also joined the ERM with a wider band.

CHAPTER 27

Business cycles

Learning outcomes

By the end of this chapter, you should be able to:

- 1 distinguish between trend growth and economic cycles around this path
- 2 discuss why business cycles occur
- 3 analyse why output gaps may fluctuate
- 4 discuss whether potential output also fluctuates
- 5 understand the role of dynamic general equilibrium models
- 6 contrast real business cycle models and New Keynesian analysis
- 7 assess whether national business cycles are now more correlated
- 8 apply these principles to UK business cycles
- 9 summarize key issues dividing the main schools of macroeconomic thought

This chapter discusses cyclical movements in output around its long-run trend. We explore theories of why cycles occur, and use this analysis to discuss the different schools of macroeconomic thought and the reasons why they differ.

For decades, politicians were accused of two policy failings. First, they pursued short-run stimulus too much in the pursuit of electorally popular high employment. The result was high inflation. Second, they could not resist boosting the economy just before elections, causing unnecessary and unhelpful swings in output.

By the late 1990s, many countries had made their central banks operationally independent of government control. This was intended both to control long-run inflation and to reduce electorally motivated cycles in

aggregate demand and output. This institutional change was necessary but not sufficient to achieve these aims. Fiscal policy could still be irresponsibly loose, and financial regulation could be dangerously weak.

Until 2007–08, it appeared that central bank independence had succeeded remarkably in delivering low inflation and lower output cycles. Then the financial crash happened. Interest rates were cut sharply and fiscal policy was loosened dramatically, both to bail out the financial system and to avert a catastrophic fall in aggregate demand, whose effects would be more prolonged the more hysteresis was a key feature of the labour market.

Central bank independence had not abolished boom and bust. Since 2009, most Western economies have had a mixture of recession and low growth, despite loose monetary policy and fiscal austerity that was more talked about than yet delivered.

These episodes illustrate many of the issues that we examine in this chapter. First, is there a business cycle? Output fluctuates a lot in the short run, but a cycle does not mean merely temporary departures from trend: it also requires a degree of regularity. Can we see it in the data? Can monetary and fiscal policy insulate economies from business cycles? If so, which should get the credit?

We also look at the international dimension. Can a country display cycles out of phase with those in its trading partners? Is globalization making business cycles more correlated across countries? If so, might a single monetary policy become increasingly appropriate?

27.1 Trend and cycle: statistics or economics?

In practice, aggregate output and productivity do not grow smoothly. In some years they grow very rapidly but in others they actually fall. Actual output fluctuates around this hypothetical **trend path**.

The **trend path of output** is the smooth path of long-run output once its short-term fluctuations are averaged out.

The **business cycle** is the short-term fluctuation of total output around its trend path.

Figure 27.1 shows a stylized picture of the business cycle. The purple curve is the steady growth in trend output over time. Actual output follows the green curve. Point A represents a slump, the bottom of a **business cycle**. At B, the economy has entered the recovery phase of the cycle. As recovery proceeds, output climbs above its trend path, reaching point C, which we call a boom. Then the economy enters a period in which output is growing less quickly than trend output, and is possibly even falling. When output is falling for at least two successive quarters, we call this a recession. Point E shows a slump, after which recovery begins and the cycle starts again.

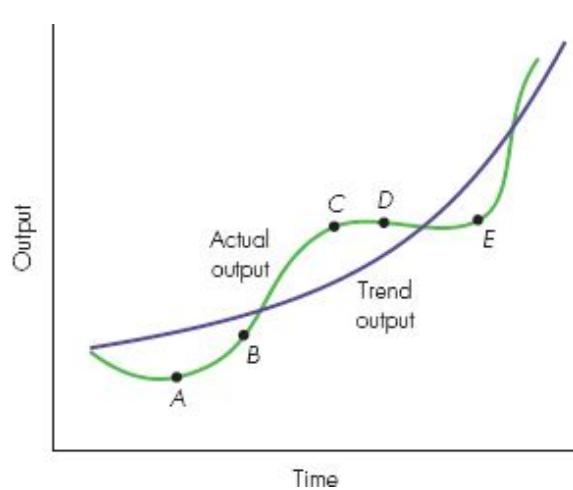


Figure 27.1 The business cycle

Trend output grows steadily over time as productive potential increases. Actual output fluctuates around this trend. Point A shows a slump, the trough of a cycle. At B recovery has begun and it continues until the peak of the cycle is reached at C. At C there is a boom. Then a period of recession follows until the next slump is reached at E. It takes roughly five years to move from one point in the cycle to an equivalent point in the next cycle, for example from A to E.

Figure 27.2 shows the annual growth of real GDP and of real output per employed worker in the UK during the period 1975–2012. The figure makes four points.

First, growth of output and productivity fluctuates in the short run. Second, although cycles are not perfectly regular, there is evidence of a pattern of slump, recovery, boom and recession, over five-six years. Third, output and output per person are *closely* correlated in the short run. Typically, output fluctuations used to precede fluctuations in productivity by about a year; since the mid-1990s this gap has fallen. Fourth, during

1995–2007 cycles became less pronounced than previously – hence the optimism that boom and bust had been defeated. The output fall of 2009 was then the worst annual fall of the post-war period. The rest of the chapter seeks to explain these facts.

Any series of points may be decomposed statistically into an average trend and fluctuations around that trend. We first assume that potential output grows smoothly. Later we ask whether potential output itself can fluctuate significantly in the short run. Thus, we start by assuming that business cycles reflect fluctuations in the [output gap](#).

Figure 27.2 shows that business cycles are too regular to be a coincidence. What causes cycles? Potential output reflects aggregate supply in the long run. First, we explore aggregate demand shocks as the source of cyclical deviations of actual output from potential output.

The [output gap](#) is the deviation of actual output from potential output.

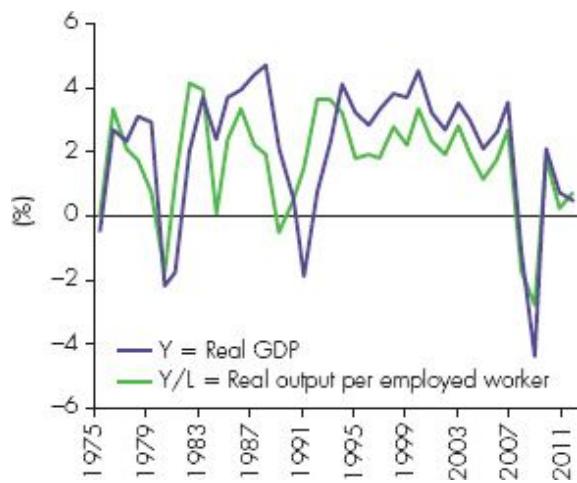


Figure 27.2 UK output and productivity growth, 1975–2009 (% pa)

We know what shifts demand: changes in export demand, in the desire to save, in expected future profits and incomes, and in monetary and fiscal policy.

If demand shocks happen to be cyclical, this causes cycles in actual output. That is not an *explanation* of the business cycle: it does not tell us why demand shocks have this cyclical pattern. One version of this approach does at least claim to be a theory.

Suppose voters, having short memories, are heavily influenced by how the economy is doing immediately prior to the election. Knowing this, the government uses monetary and fiscal policy to manipulate aggregate demand. Policy is tight just after a government is elected, creating a slump and spare capacity. As the next election date approaches, expansionary policy can then create unsustainably rapid growth by eliminating spare capacity. Voters misinterpret this as permanently faster growth of potential output and gratefully re-elect the government.

A **political business cycle** arises if politicians manipulate the economy for electoral advantage.

This theory provides a reason for fluctuations and suggests why **political business cycles** tend to last about five years—that is about the period between successive elections. The theory probably contains a grain of truth. On the other hand, it supposes that voters are naïve. Voters are not always so short-sighted. In 1997 the Major government lost the election despite fast output growth: voters thought Labour could do even better. In the US, President Obama was re-elected in 2012 despite low growth since 2009.

Recent institutional changes to improve the credibility of policy—particularly central bank independence—reduce the scope for political business cycles. Having discussed political causes of cycles, we now concentrate on economic causes.

27.2 Theories of the business cycle

Fluctuations in export demand might cause cycles. One country's exports are another country's imports, and these imports will fluctuate only if foreign income fluctuates. International trade helps explain how cycles get transmitted from one country to another, but we require a theory of domestic business cycles to initiate the process.

Sluggish adjustment is necessary but not sufficient to generate cycles caused by demand shocks. It is necessary because rapid adjustment would quickly eliminate output gaps and restore output to potential output. It is not sufficient because sluggishness only explains why the return to potential output takes time. An oil tanker moves sluggishly but it does not oscillate its way into port. Cycles require a mechanism by which

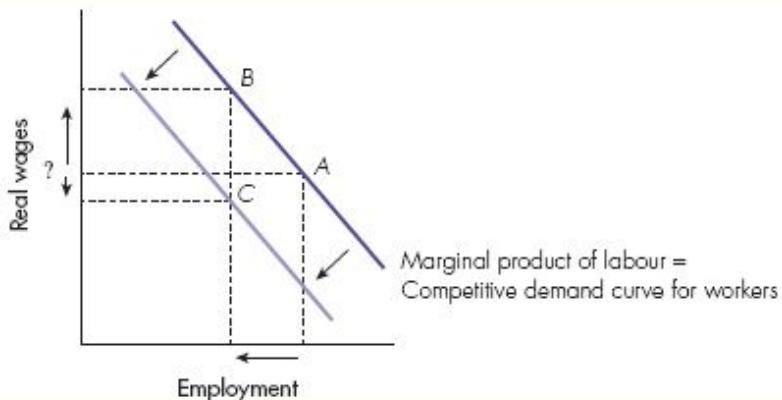
deviations in one direction then set up forces that cause output to overshoot potential output on its return.

Having ruled out the government, a theory of domestic cycles must be based on consumption or investment spending. Investment spending is the most likely candidate, since it is more likely to take time to assess and adjust. Firms do not rush into major and irreversible investment projects, nor are new factories built overnight.

ACTIVITY 27.1

THE CYCLICAL BEHAVIOUR OF WAGES

In a recession, firms employ fewer workers. A competitive firm would pay workers the real value of their marginal product. Given a diminishing marginal product of labour, cutting back on workers should raise labour's marginal product. Fewer workers have the same capital as before to work with. Real wages should rise in a slump, as the economy moves from point *B* to *A* in the figure below. But in practice, real wages don't rise, they fall. This is the *real-wage puzzle* over the business cycle.



To explain the real-wage puzzle, we have to explain why the demand for labour shifts down so the economy moves in a slump from point *A* to *C* not from *A* to *B*. One answer is that most of the economy faces imperfect competition. When demand for output falls, profit margins are reduced, so the marginal revenue product of labour falls too, reducing the demand for labour. Profit margins are not constant over the business cycle.

Second, at lower output levels, firms may scrap old factories and old machines that were only just sustainable provided output remained

high, reducing the total stock, causing a downward shift in labour demand, even under perfect competition in the output markets.

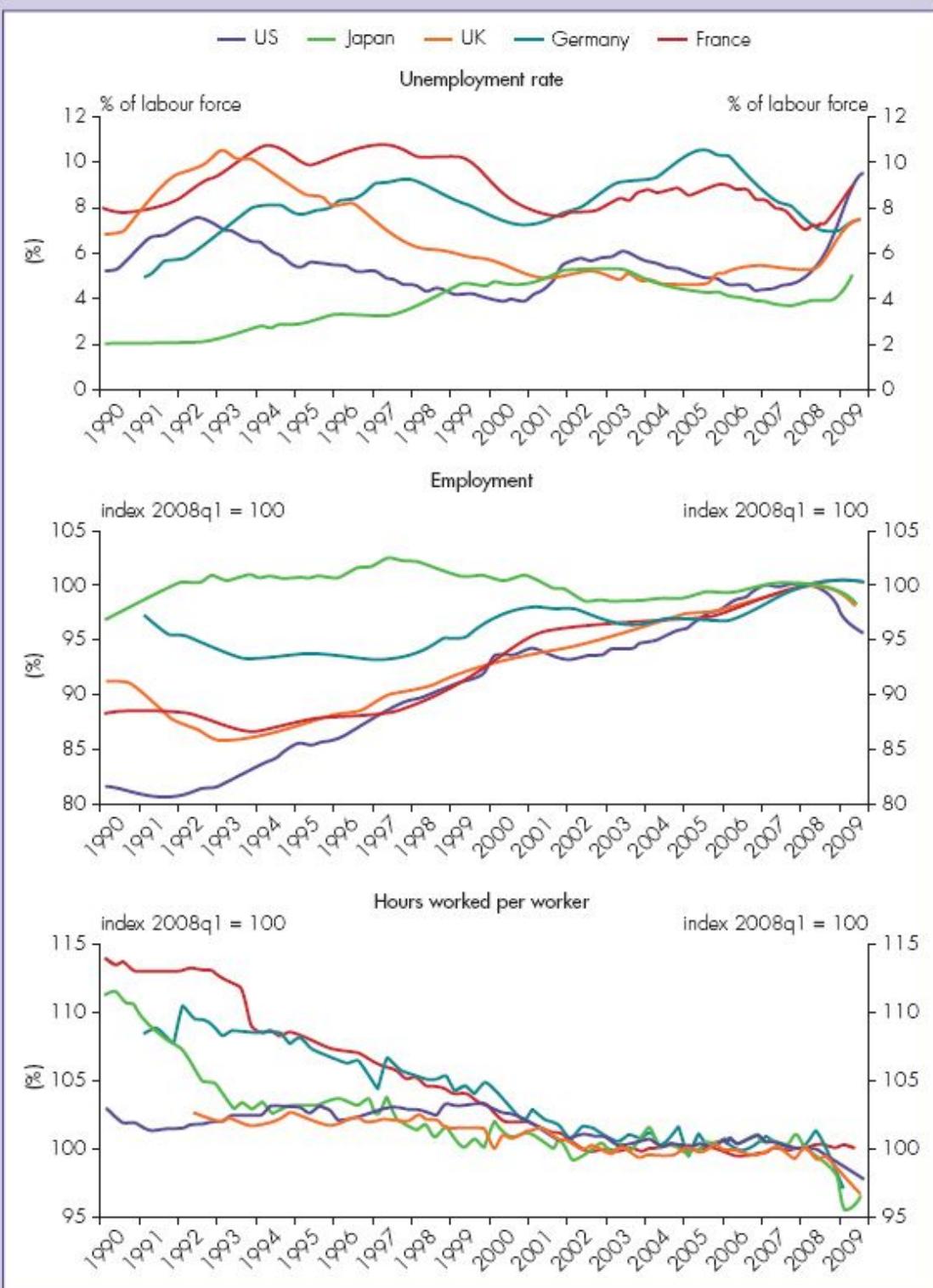
Third, trade union power may be weakened by a recession and the prospect of job cuts, leading to a lower wage settlement. In effect, unions cut wages in order to preserve employment for their members.

Fourth, some of the adjustment to recession may come via less overtime and shorter working weeks. This makes sense if it is cheaper to adjust hours of work than to hire and fire workers. Shorter working hours mean both lower output per worker and lower take-home pay per worker. The cost of firing, and then subsequently rehiring, means that firms may actually hoard labour in a recession, keeping on workers that they don't fully need today in anticipation of better times to come tomorrow.

Nor do recessions originate only from adverse demand shocks in the output market. Suppose there is a temporary adverse supply shock – either a surge in raw materials prices or a temporary reduction in labour productivity. In either case, the demand for labour is reduced, as in the figure above.

Shifts in labour supply may further complicate the analysis. When times are tough and wages low, you don't sacrifice much by taking time off; lifetime earnings can be rebuilt when conditions are easier. So recessions, caused by temporarily low productivity, make firms offer temporarily low wages, and households temporarily reduce their labour supply. We can get low employment *and* low wages.

The next set of figures, taken from the OECD's *Economic Outlook* of November 2009, confirm that this phenomenon is pervasive. Hours worked per worker (bottom panel) fell sharply in 2008, just before employment turned down (middle panel) and unemployment turned up (top panel).



Questions

- (a) If the labour demand curve does not shift, should lower employment be associated with higher or lower real wages?

- (b) In which direction might labour demand curves shift during a recession? What would you then expect the correlation to be between changes in wages and changes in employment?
- (c) Would it matter whether the recession was caused by an adverse demand shock or an adverse supply shock?
- (d) When it is expensive to hire and fire workers, how are firms likely to react to a recession that is perceived as temporary?
- (e) If firms then become pessimistic about the persistence of the recession, what is likely to happen to their demand for labour?

To check your answers to these questions, go to page 682.

The multiplier-accelerator model of the business cycle

This model distinguishes between the consequences and the causes of a change in investment spending. The consequence is straightforward. In the simplest Keynesian model, higher investment leads to a larger rise in income and output in the short run. Higher investment not only adds directly to aggregate demand but, by increasing incomes, it also adds indirectly to consumption demand. Chapters 16 and 17 examined the multiplier effect on output.

What about the cause of a change in investment spending? Firms invest when their existing capital stock is smaller than the capital stock they would like to hold. When firms are holding the optimal capital stock, the marginal cost of another unit of capital just equals its marginal benefit—the present value of future operating profits to which it is expected to give rise over its lifetime. This present value can be increased either by a fall in the interest rate at which the stream of expected future profits is discounted or by an increase in the future profits expected.

Thus far we have focused on the role of changing interest rates in changes in investment demand. However, although nominal interest rates change a lot, real interest rates change a lot less. The simplest way to calculate the present value of a new capital good is to assess the likely stream of *real* operating profits (by valuing future profits at *constant prices*) and then to discount them at the *real* interest rate.

In practice, changes in interest rates may *not* be the most important source of changes in investment spending. Almost certainly, changes in

expectations about future profits are more important. The dotcom bubble collapsed not because of high real interest rates but because people realized they had been too optimistic about the future profits to be made.

More generally, if real interest rates and real wages change slowly, the main source of short-term changes in beliefs about future profits is beliefs about future levels of sales and output. Other things equal, higher expected future output raises expected future profits and raises demand for investment in new capacity. This is the insight of the [accelerator model of investment](#).

The accelerator is only a simplification. A complete model of investment would examine changes in expected future profits and changes in (real) interest rates. Even so, many empirical studies confirm that the accelerator is a useful simplification.

How firms respond to changes in output depends on two things: first, the extent to which firms believe that current output growth will be sustained in the future; second, the cost of quickly adjusting investment plans, capital installation and the production techniques thus embodied. The more costly it is to adjust *quickly*, the more firms spread investment over a longer period.

The [accelerator model of investment](#) assumes that firms guess future output and profits by extrapolating past output growth. Constant output growth leads to a constant level of investment. It takes *accelerating* output growth to *raise* desired investment.

This simple multiplier-accelerator model can lead to a business cycle. In Table 27.1 we make two specific assumptions, although the argument holds much more generally. First, we assume that the value of the multiplier is 2. An extra unit of investment raises income and output by 2 units. Second, we assume that current investment responds to the growth in output *last* period. If last period's income grew by 2 units, we assume that firms raise current investment by 1 unit.

Table 27.1 The multiplier-accelerator model of the business cycle

Period	Change in last period's output ($\gamma_{t-1} - \gamma_{t-2}$)	Investment I_t	Output γ_t
$t = 1$	0	10	100
$t = 2$	0	10	120
$t = 3$	20	20	140
$t = 4$	20	20	140

$t = 5$	0	10	120
$t = 6$	-20	0	100
$t = 7$	-20	0	100
$t = 8$	0	10	120
$t = 9$	20	20	140

In period 1, the economy is in equilibrium with output $Y_I = 100$. Since output is constant, last period's output change was zero. Investment $I_1 = 10$, which we can think of as the investment needed to offset depreciation and keep the capital stock intact.

Suppose in period 2 that some component of aggregate demand rises by 20 units. Output increases from 100 to 120. Since we have assumed that a growth of 2 units in the previous period's output leads to a 1-unit increase in current investment, the table shows that in period 3 there is a 10-unit increase in investment in response to the 20-unit output increase during the previous period. Since the assumed value of the multiplier is 2, the 10-unit increase in investment in period 3 leads to a further increase of 20 units in output, which increases from 120 to 140.

In period 4 investment remains at 20 since the output growth in the previous period was 20. Thus output in period 4 remains at 140. But in period 5 investment reverts to its original level of 10, since there was no output growth in the previous period. This fall of 10 units in investment leads to a multiplied fall of 20 units in output in period 5. In turn, this induces a further fall of 10 units of investment in period 6 and a further fall of 20 units in output.

Since the rate of output change is not accelerating, investment in period 7 remains at its period 6 level. Output is stabilized at 100 in period 7. With no output change in the previous period, investment in period 8 returns to 10 units and the multiplier implies that output rises to 120. In period 9 the 20-unit increase in output in the previous period increases investment from 10 to 20 units and the cycle begins all over again.

The multiplier-accelerator model explains business cycles by the dynamic interaction of consumption and investment demand. The insight of the model is that it takes accelerating output growth to increase investment. Once output growth stabilizes, so does investment. In the following period, investment must fall, since output growth has been reduced. The

economy moves into a period of recession, but once the rate of output fall stops accelerating, investment starts to pick up again.

This simple model is not the definitive model of a business cycle. If output keeps cycling, surely firms stop extrapolating past output growth to form assessments of future profits? Firms, like economists, recognize that there is a business cycle. The less investment decisions respond to the most recent change in past output, the less pronounced will be the cycle.

MATHS 27.1

THE MULTIPLIER-ACCELERATOR MODEL OF CYCLES

Suppose I denotes current investment, I_{-1} denotes investment in the previous period, Y denotes output and ΔY denotes $(Y - Y_{-j})$, the increase in output between last period and the current period. Output Y is related to current investment I by the multiplier $Y = I/(1 - c)$, where c is the marginal propensity to consume. Investment depends on output growth, so $I = a \Delta Y$. Hence,

$$I = a \Delta Y = [a/(1 - c)][I - I_{-1}]$$

Hence,

$$I = -[a/[1 - c - a]]I_{-1}$$

This equation is of the general form $I = bI_{-1}$. If b is a positive fraction, I is always smaller than the period before and gradually converges on zero. If b exceeds unity, I gets larger and larger for ever. Negative values of b imply I becomes negative every second period, either converging to zero or becoming ever larger. None of this generates things like business cycles.

Cycles emerge, however, with small changes to these formulae. Table 27.1 offers one example. Here is another. Suppose the consumption function depends not on current income but on previous period income so that $C = A + cY_{-1}$ and current investment depends on output growth in the previous period, so that $I = a[Y_{-1} - Y_{-2}]$. Since $Y = C + I$ in this simple economy,

$$Y = A + cY_{-1} + a[Y_{-1} - Y_{-2}] \quad (1)$$

If the economy is in long-run equilibrium, output is constant, the final term is zero, and equilibrium output Y^* is given by $Y^* = A/(1 - c)$. Using y to denote $Y - Y^*$, the deviation of output from its long-run level, we can subtract Y^* from both sides of equation (1) to yield

$$y = cy_{-1} + a[y_{-1} - y_{-2}] \quad (2)$$

Depending on the values of c and a , equation (2) can yield constant cycles, damped cycles that gradually get smaller and smaller or explosive cycles that get larger and larger. When $c = a$, we simply get

$$(y - y_{-1}) = -(y_{-1} - y_{-2})$$

so that positive and negative growth of similar size alternate for ever.

Ceilings and floors

The multiplier-accelerator model can generate cycles even without any physical limits on the extent of fluctuations. Cycles are even more likely when we recognize the limits imposed by supply and demand. Aggregate supply provides a *ceiling* in practice. Although it is possible temporarily to meet high aggregate demand by working overtime and running down stocks of finished goods, output cannot expand indefinitely.

This tends to slow down growth as the economy reaches a boom. Having overstretched itself, the economy has to bounce back off the ceiling and begin a downturn. Conversely, there is a *floor*, below which aggregate demand cannot fall. Gross investment (including replacement investment) cannot be negative unless, for the economy as a whole, machines are unbolted and shipped abroad for sale to foreigners. Falling investment is an important component of a downswing, but investment cannot fall indefinitely, whatever our model of investment behaviour.

Fluctuations in stockbuilding

Having examined investment in fixed capital, we now look at inventory investment in working capital. Firms hold stocks of goods despite the cost; namely, the interest payments on the funds tied up in producing the goods for which no revenue from sales has yet been received. What is the corresponding benefit of holding stocks? If output could be instantly and

costlessly varied, it would always be possible to meet sales and demand by varying current production. Holding stocks makes sense because it is expensive to adjust production *quickly*. Output expansion may involve heavy overtime payments and costs of recruiting new workers. Cutting output may involve expensive redundancy payments. Holding stocks allows firms to meet short-term fluctuations in demand without incurring the expense of short-run fluctuations in output.

How do firms respond to a fall in aggregate demand? Since rapid output adjustment is expensive, in the short run firms undertake the adjustments that can be made more cheaply. They reduce hours of overtime and possibly even move on to short-time working. If demand has fallen substantially, this still leaves firms producing a larger output than they can sell. Firms build up stocks of unsold finished output.

If aggregate demand remains low, firms gradually reduce their workforce. Natural wastage occurs (departing workers are not replaced). It also becomes cheaper to sack some workers than to meet the interest payments on holding ever-larger volumes of stock. Once aggregate demand recovers, firms are still holding all the extra stocks built up during recession. Only by increasing output *more slowly* than the increase in aggregate demand can firms eventually sell off these stocks and get back to their long-run equilibrium position.

Costs of employment adjustment explain both the pattern of inventories over the business cycle and the pattern of labour productivity in Figure 27.2. Output per worker rises in a boom and falls in a slump. In other words, output adjusts more quickly than employment. This is what we expect, given the costs of adjusting employment rapidly.

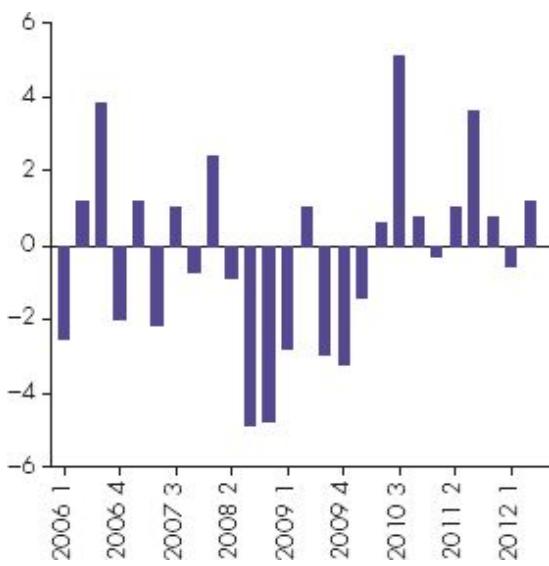


Figure 27.3 UK stockbuilding, 2006–2012 (£bn)

Source: Bank of England, Inflation Report.

A fall in demand is met initially by cutting hours and increasing stocks. With a shorter working week, there is a fall in output per worker. Only as the recession intensifies do firms undertake the costlier process of sacking workers and restoring hours to their normal level. Conversely, a boom is the time when output and overtime are high and productivity per worker peaks.

Figure 27.3 confirms this clearly in response to the fall in aggregate demand that began in 2008. During 2006–07, the level of stocks fluctuated from quarter to quarter without any particular trend. In the first quarter of 2008, aggregate demand fell and firms were left with unsold goods. Their stocks rose unexpectedly. Foreseeing that demand would be weak for some time, firms began to cut back production, reducing stocks of work in progress. When production had fallen more than demand, stocks of unsold finished goods also fell. Figure 27.3 shows substantial destocking during the rest of 2008 and early 2009. Once production had fallen more than demand, stocks started to increase again.

Competitiveness

Chapter 24 identified another potential mechanism that could generate cycles. An economy on a fixed exchange rate experiences a downward

domestic demand shock. Interest rates, fixed at world levels to peg the exchange rate, cannot be used to restore aggregate demand.

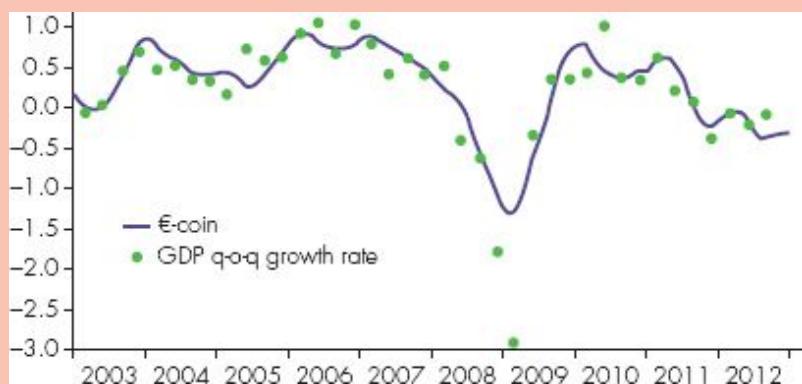
Recession eventually bids down wages and prices, thus raising competitiveness and restoring internal balance by raising the demand for net exports. However this is not external balance, since net exports are now positive. With a current account surplus, the country gets richer, and additional wealth gradually boosts consumption demand. The economy now has a boom, which bids up prices and reduces competitiveness. Long-run equilibrium is restored when the current account falls back to zero.

This is a proper story about cycles. Output gaps induce changes in the price level that restore internal balance only by destroying external balance. This sets off a movement in the opposite direction that gradually reverses all these effects. Adjustment entails necessary overshooting of the final equilibrium.

CASE 27.1

EUROZONE BUSINESS CYCLES

EuroCOIN, the monthly coincident indicator of the Eurozone business cycle, is published by the Centre for Economic Policy Research based in London. The figure shows values of the indicator (in purple) and the quarterly growth rate of Eurozone GDP (in green) during 2003–12.



€-Coin and euro-area GDP

Source: <http://eurocoin.cepr.org/>.

Why not use changes in GDP itself to measure the business cycle? Mainly because initial estimates of GDP are unreliable and the data are often revised a lot as time elapses. The *EuroCOIN* indicator not

only estimates the cyclical component of GDP more accurately but is also available monthly, whereas GDP estimates appear only quarterly. By examining past correlations of GDP growth with data that do appear monthly, the indicator provides a more frequent and more reliable picture of the Eurozone business cycle—helpful information for the monthly meetings of the European Central Bank at which interest rate decisions are made.

The figure shows that, like other independent central banks, the European Central Bank had been fairly successful up to 2007 in stabilizing output. There was not much of a business cycle. The monthly *EuroCOIN* indicator shows a slowdown beginning to happen even during 2007 and then rapidly during 2008. The indicator dates early 2009 as the cyclical bottom for output growth, even though the growth indicator did not climb above zero until later in 2009, and did not deliver two successive quarters of positive growth—the official definition of the end of the recession—until the end of 2009. By 2012 the Eurozone was back in recession.

Source: <http://www.cepr.org/Data/eurocoin>.

27.3 Real business cycles

So far our analysis of business cycles focuses on demand shocks and cyclical movements in output gaps. This is compatible with our earlier analysis of sluggish wage adjustment in the short run. This view of cycles is consistent with a model that is Keynesian in the short run, but classical or monetarist in the long run.

Real business cycle theories explain cycles as fluctuations in potential output itself.

Not all economists share our assessment of how the economy works. In particular, there is an influential school, known as the New Classical economists, whose intellectual leader is the Nobel Laureate Robert Lucas of the University of Chicago. A key assumption of the New Classical school is that all markets clear almost instantaneously. Effectively, output is almost always at its full-employment level.¹

Proponents of the theory argue that macroeconomics should base theories of firms and households in a microeconomic analysis of choice between

the present and the future. For example, this approach would view each household as making a plan to supply labour and demand goods both now and in the future in such a way that lifetime spending was financed out of lifetime income plus any initial assets. Such plans would then be aggregated to get total consumption spending and total labour supply. An equivalently complex story would apply to firms and investment.

One implication of this approach is that it is no longer helpful to distinguish between supply and demand. If labour supply and consumption demand are part of the same household decision, things that induce the household to change its consumption demand also induce it to change its labour supply.

For this reason, real business cycle theorists simply discuss what happens to actual output, which reflects both supply and demand and, by assumption, equates the two at potential output. In this view, the economy is then bombarded with shocks (for example, breakthroughs in technology, changes in government policy, changes in oil prices), which alter these complicated plans and give rise to equilibrium behaviour that looks like a business cycle.

Why is this approach called the *real* business cycle approach? In the classical model, nominal money only affects other nominal variables. Output and employment depend only on real variables. Since real business cycle theorists believe in the classical model, they take it for granted that the source of business cycles must be in real shocks. Fancy dynamics can then explain why shocks last and have convoluted effects.

Intertemporal substitution: a key to persistence

Real business cycle theories need to combine rapid market adjustment to equilibrium with sluggish behaviour of aggregate output over the business cycle. Intertemporal substitution means making trade-offs over time, postponing or bringing forward actions in the sophisticated long-run plans of households and firms. This behaviour can cause effects to persist and look like part of a business cycle.

Suppose the productivity genie visits while we are all asleep. When we wake up, our productivity has doubled, but only for a year. We know that by next year our productivity will have returned to normal. We face a temporary productivity shock, a blip in our technology. What should we do?

We are definitely wealthier after the genie's visit. We are pleased it happened. We could simply behave as before, working just as hard and investing just as much. In that case, our extra productivity would make extra output this year, but it is output that we would blow entirely on consumption this year. We would get little extra utility out of the hundredth bottle of champagne, and we would be making no provision for the future. There must be a better way.

We could put in a temporary spurt of extra work while we are super-productive, but in itself that would only exacerbate the problem: even more champagne today, still nothing extra for tomorrow. In fact, because leisure is a luxury and because we are better off than before, we may feel like taking it easy and doing less work.

We need a way of transferring some of our windfall benefit into future consumption. The solution is investment. A sharp rise in the share of output going to investment will provide more capital for the future, thereby allowing higher future consumption even after our productivity bonus has evaporated. Once we get to the future, being then richer than we would have been without the genie, we may in consequence work less hard than we would have done, since leisure is a luxury.

The point of this example is to show that even a temporary shock can have effects that persist well into the future. Persistence occurs both through investment (in human as well as physical capital) and through intertemporal labour substitution—deciding when in one's life to put in the effort.

Real business cycle theories continue to be developed fully. Apart from optimism about the speed of adjustment, they have been criticized on two grounds. First, they are usually theories of persistence not cycles. Shocks have long-drawn-out effects, but rarely are these cyclical. To 'explain' business cycles, so far real business cycle theorists have had to assume a cyclical pattern to the shocks themselves. The theory is therefore incomplete.

Second, and related, since the most widely researched example involves shocks to technology, a cyclical pattern of shocks implies that in some years technical knowledge actually diminishes: we forget how to do things. Not just once, but regularly every few years. This may be a bit hard to swallow.

However, this can be given a more plausible interpretation. In the dotcom bubble of the late 1990s, investors made extravagant projections about

future productivity growth and associated profits from the new technologies. By 2000 evidence was accumulating that previous estimates, necessarily guesses in a new situation, were too optimistic. In 2001 investment collapsed, particularly in the US where dotcom optimism had been greatest.

Thus, the adverse shock was not a fall in existing technology—which is indeed implausible—but in estimates of future technology, which affects current behaviour since firms, households and governments all make long-term plans.

Policy implications

Research on real business cycles has one vital message for macroeconomic policy. If the theory is right, it destroys the case for trying to stabilize output over the business cycle. Fluctuations in output are fluctuations in an *equilibrium* output that efficiently reconciles people's desires.

For example, in the parable of the genie, the induced effects on investment, labour supply, output and consumption implement people's preferred way to take advantage of the beneficial opportunity. Trying to prevent these ripples is misguided policy.

Although important, this caveat undermines the case for stabilization policy only if we buy totally the assumptions of complete and instant market-clearing and the absence of any externalities. For most economists these assumptions are too extreme to reflect the real world, which continues to exhibit Keynesian features in the short run. Valid reasons for stabilization policy then remain.

Even so, real business cycle theories force us all to acknowledge that there is no reason why potential output should grow as smoothly as trend output. The latter is a statistical artefact whose construction, averaging, forces it to be smooth.

Credit constraints and aggregate supply

The financial crash provided a huge adverse shock to aggregate demand, as people watched their wealth evaporate. But it would be wrong to assume the crash had no direct effect on aggregate supply. Firms need to borrow to finance the costs of production before this output can be sold.

When banks became insolvent, or people feared that banks were close to being insolvent, this had two effects. First, banks had inadequate reserves to take their normal business risks. Second, banks and all other financial market participants suddenly raised their estimate of the likely riskiness of borrowers. The result, as we saw in Chapter 18, was that interest rate spreads became huge and banks stopped lending almost completely.

This meant that many businesses found it impossible to finance production and had to cut back. Aggregate supply fell, independently of what was happening to aggregate demand.

27.4 Supply-side effects of the financial crash

We began this chapter by exploring mechanisms through which we might generate cycles in aggregate demand. These include the multiplier-accelerator model, the effects of stockbuilding, the consequences of fixed exchange rates, political cycles in the policy stimulus, and simple effects of ceilings and floors. Within such frameworks, fluctuations in aggregate demand lead to similar fluctuations in output gaps, since the trend behaviour of potential output is unaffected.

Fluctuations in aggregate demand could be largely offset by an independent central bank with perfect foresight. Eliminating fluctuations in output gaps would eliminate a key source of inflationary pressure and help stabilize inflation. The fact that there is substantial evidence that business cycles were much less marked during 1995–2007, when independent central banks were explicitly asked to conduct this task, is *prima facie* evidence that demand fluctuations had been the most frequent source of business cycle fluctuations. Since central banks do not have perfect foresight, and since interest rate changes take time to affect aggregate demand, even excellent central banks could never have been expected to eliminate cycles completely.

Demand is very important, but not the whole story, for four reasons. First, real business cycle theories provide a healthy antidote to an exclusive focus on demand. Sometimes, supply-side factors will cause uneven growth of potential output, and may even reduce potential output for a while. Nobody promised that technological progress would occur at an even rate.

Second, hysteresis is important, especially for large shocks. Chapter 23 explored how temporary shocks to demand can have lingering, even permanent, effects on supply, through induced effects on the capital stock, skill base and union power. Third, credit rationing by banks and other lenders can directly curtail the ability of firms to finance production.

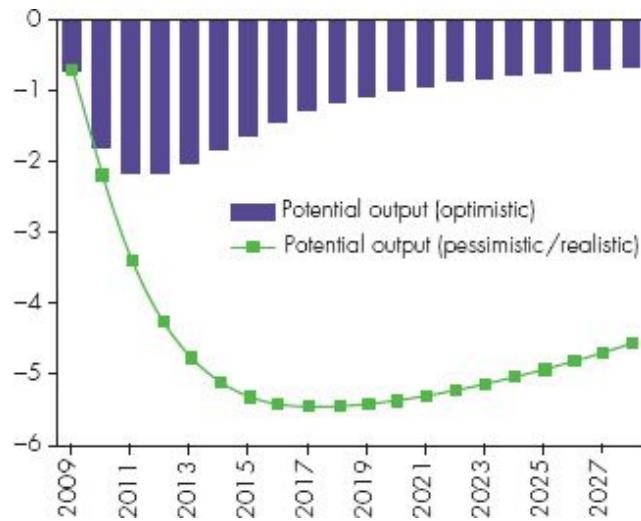


Figure 27.4 Possible impact of crash on potential output

Source: EU Commission, European Economy, 2009. © 2010 The World Bank Group. All rights reserved.

Research for the EU Commission has tried to model how large these effects might be.² Since we do not have enough evidence to give a definitive estimate, the Commission economists have produced an ‘optimistic’ and ‘pessimistic’ scenario, shown in Figure 27.4. In the optimistic scenario, the crisis leads to a fall in EU potential output by about 2 per cent in 2011, after which potential output gradually reverts to the path it would anyway have followed. However, even in the optimistic scenario, the financial crisis and subsequent recession is still casting a shadow on potential output as late as 2025. Since output cannot remain above potential output indefinitely, anything that reduces potential output for 15 years entails sacrificing considerable output cumulatively, even after the immediate crisis has been overcome.

In the pessimistic scenario, the induced effects on potential output are much, much larger and last much longer. Potential output gets worse until 2017 and is still 5 per cent worse than it would have been as late as 2025.

From the perspective of a practical politician, this is as bad as a permanent reduction in potential output.

In order fully to understand business cycles, and hence to assess the consequences of the financial crash, we need to be quite sophisticated. Aggregate demand fluctuates (unless stabilized by monetary policy), but this should probably be superimposed on a path of potential output that is also capable of fluctuations. Although it is an empirical question which is larger, in normal circumstances the possible fluctuations in aggregate demand are probably more important.

In more extreme circumstances, and particularly where a financial crash is involved, aggregate supply is also capable of falling sharply for two reasons: the credit impact on the ability to finance production, and the various hysteresis effects that demand falls then induce as supply falls.

CONCEPT 27.1

DYNAMIC STOCHASTIC GENERAL EQUILIBRIUM MODELS

Now for a brief glimpse of the coal face, at which many of the top academic research economists spend their time. This approach seeks to integrate microeconomics and macroeconomics to produce a single theory of everything in macroeconomics. It begins with microeconomic descriptions of the preferences of individuals (their tastes over consumption and leisure), the technology available to firms (how combinations of inputs make output) and the institutions characterizing the economy (the solvency constraints on households, firms and governments, and the nature of the political process and policy making).

By working with representative agents in each sector, it is possible to build up aggregate behaviour and hence *general equilibrium* in all markets simultaneously. By specifying the processes driving random shock that hit tastes and technology, this becomes a random or *stochastic* model, in which agents form guesses about their likely futures. It is a *dynamic* model because the present and the future are connected. Consumers think about long-run income in making lifetime work and spending plans, governments and firms can also borrow to cover investments made today that generate future benefits.

These dynamic stochastic general equilibrium (DSGE) models essentially fall into two groups. The first are the real business cycle models we discussed earlier in the chapter, which focus on how real shocks (to tastes or technologies) generate short-run fluctuations within a framework of overall long-run growth. Real business cycle models have no role for stickiness of nominal prices or wages, and no role for benevolent government intervention to help the economy back to potential output. The second group of DSGE models are called **New Keynesian** models.

New Keynesians provide rigorous microeconomic foundations for macroeconomics with temporary price rigidity.

In earlier chapters, we developed Keynesian analysis from a series of plausible but ad hoc assumptions. The New Keynesian approach is a modern response to the claim that Keynesian economics is not properly grounded in serious microeconomics. The essential trick of New Keynesian economics is to take the DSGE model but to replace the assumption of price flexibility with the assumption that monopolistically competitive firms set prices that cannot be instantly and costlessly adjusted. This reintroduces nominal rigidities in the short run. The complicated DSGE model is then used to churn out the implications, which of course turn out to be very similar to the implications of the analysis we have adopted in the preceding chapters. In particular, demand fluctuations can generate short-run cycles. A stabilizing role for monetary policy is restored.

The elite end of the economics profession would like to show that their subject is just as scientific as physics or chemistry. Deriving behaviour from first principles is appealing, and stochastic dynamic general equilibrium models look rigorous. But there is a problem. Recall our distinction between micro and macro in Chapter 1 . Our brains have chips that are too small to process everything all at once: that is why the distinction between micro and macro remains meaningful.

DSGE models appear to generate understandable solutions to complex problems, but do so largely by falling back on a different, but equally strong, simplification, that of the individual agent behaviour that is easily aggregated to describe total behaviour. Once the mix of individuals (or firms) becomes messy, it becomes

impossible to aggregate easily. So DSGE models—whether real business cycle or New Keynesian—appear rigorous, but behaviour has been simplified in order to make it easily analysable within this complex setting. In that sense, it is just another shortcut, with no more (and no less) legitimacy than the shortcuts we have adopted to present the macroeconomic analysis of the previous chapters.

The ultimate test is which model policy makers used to guide their actions at the height of the financial crisis. They found insights from them all: the credit shocks of the real business cycle model, the nominal shocks in the New Keynesian framework. However, in a crisis, simple robust tools often work best. Those equipped only with the preceding macro chapters would have done just as well in understanding the evolving crisis as those seeking to interpret it through more elegant, but more complicated, DSGE models.

27.5 An international business cycle?

National politicians want all the credit when output is high but produce a cast-iron alibi when the economy turns sour. They say domestic difficulties were caused by a world recession. How good is their alibi?

Figure 27.5 plots data during 1996–2009 for the US, Japan, the UK, Germany, France and Italy. Although there were differences in the 1990s—for example, Japan continued to stagnate—it is striking how similar output growth has been across countries since 2000. Largely this reflects the fact that both the dotcom bust and the financial crisis were global events. In addition, independent central banks have pursued rather similar monetary policies, eliminating one important source of differences in national policy shocks.

These patterns warn us how interdependent the leading countries have become in the modern world. Economies are becoming more open. In product markets, protectionist policies are being removed, through global institutions like the World Trade Organization and through regional integration, as in the creation of a Single European Market.

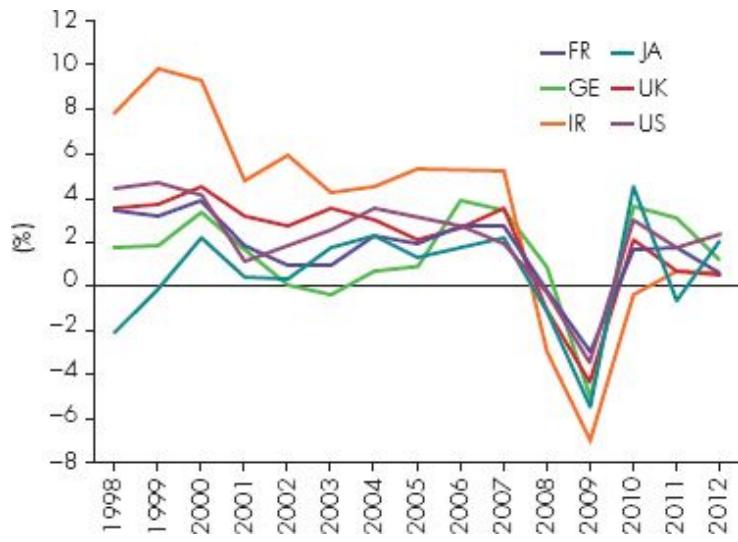


Figure 27.5 Recent business cycles, 1998–2012 (GDP growth, % pa)

Source: OECD, Economic Outlook.

Improvements in transport and telecommunications also favour greater integration of product markets. When R&D costs are large, producers need a global market if they are to recover their overheads. Product market integration provides an international transmission mechanism through exports and imports. Increasingly, we have a global financial market. Closer financial integration increases the likelihood that different countries pursue similar monetary policies.

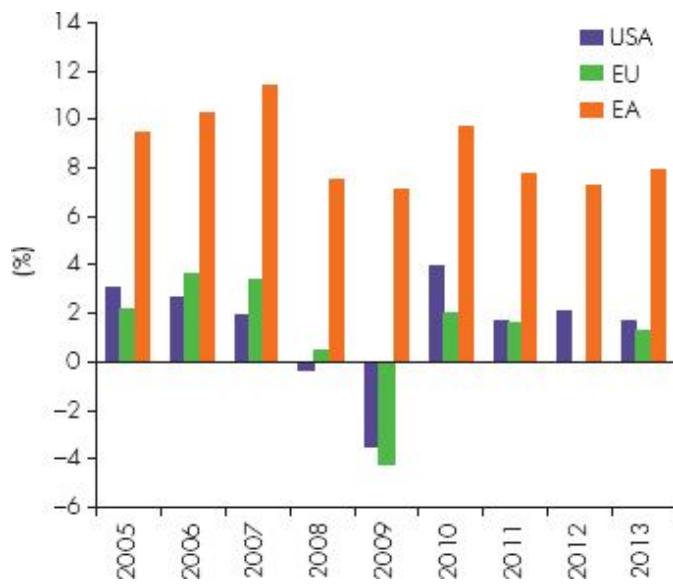


Figure 27.6 Annual growth rates, US, EU, Emerging Asia, 2005–13 (%)

Source: IMF, World Economic Outlook, 2012.

The business cycle is transmitted from one country to another not just through private sector decisions about imports and exports (and induced effects on labour supply, investment and consumption), but also, sometimes, through induced changes in the economic policy of other governments.

The considerable integration of advanced economies prompts a second question: are the leading emerging market economies, particularly China and India, on the way to similar integration with their longer-established partners? Figure 27.6 shows the most recent part of the answer. Prior to the crisis, Asian economies had been growing much more quickly as they caught up with the OECD. The cyclical pattern in response to the global crisis was very similar to OECD countries, albeit from a higher baseline rate of growth.

Chinese and Indian markets were not sheltered from the effects of the US sub-prime mortgage crisis. Either the financial impact was felt directly, because Asian lenders had invested in companies such as Lehman Brothers or been induced to buy securitized products that were subsequently exposed as worthless, or the impact was a second-round effect – as Western economies contracted, exports from India and China suffered. Whatever the channel, Figure 27.6 confirms that most of the important parts of the world are now sufficiently connected that national policy makers cannot escape unscathed.

The key message, therefore, is that some shocks are experienced in common, but others are not. For most countries, globalization increases the probability that the shocks with which its policy makers have to deal are not emanating from the domestic economy but from abroad. Most of the time, only countries as large as the US, China and soon India will be big enough, acting alone, to have a major effect on the rest of the global economy.

27.6 Overview of schools of macroeconomic thinking

We have come a long way in our discussion of macroeconomics. We have slowly built up an analysis of how the economy works, and studied the effects of government policy in both the short run and the long run. It is a good time to take stock of the main competing views of macroeconomics and their implications for government policy.

We begin by highlighting the major issues on which there is important disagreement. Different assumptions lead to different conclusions. We focus on four issues: the speed with which markets clear, whether or not equilibrium is unique, how expectations are formed, and the relative importance of the short run and the long run.

Against this background we then describe the four most prominent schools of macroeconomics in recent years. We encourage you to view these competing positions not as unrelated and contradictory beliefs, but as the outcome of adopting slightly different positions within the spectrum of possible views.

Areas of disagreement

Why do economists disagree at all? Surely, by looking carefully at the evidence we can say which views are correct and which are inconsistent with the facts? Unlike some of the physical sciences, economists can rarely undertake controlled laboratory experiments. In practice, we have to try to unscramble historical data to infer how the economy works.

Empirical research in economics does not always offer clear-cut answers. We live in a world that is constantly evolving. Even if we had a good estimate of the empirical magnitudes in the demand for money equation during 1950–70, would it be relevant after 2015, with mobile phones and Internet shopping? When behaviour is changing, although we get new data, the relevance of old data becomes weaker.

Taking a different example, current behaviour is heavily influenced by expectations of the future. The spending decisions of firms and households depend critically on today's expectations of future incomes and profits. Purchases and sales let us measure actual spending, but we have no equivalent data on current expectations. Suppose a sharp rise in income and output is *not* preceded by a sharp increase in consumption and investment spending: do we conclude that nobody had previously expected income and output to rise, or that the rise was foreseen but had little effect on consumption and investment decisions? Different schools of economists look at the same data but view them differently.

Empirical economists do the best they can. In some cases their research is rather persuasive and their conclusions are accepted. Few people dispute that consumption and money demand are affected by current income. In other cases, empirical research is less conclusive. Although economists agree about many aspects of positive economics, some disagreements inevitably remain. We pick out key disagreements, not mere quibbles about points of detail. They fundamentally affect one's view of the world and the policy decisions one is likely to support.

Market clearing

A market clears when desired supply equals desired demand. Whether, and if so, how quickly all markets clear is a key issue in macroeconomics. At the one extreme, the classical analysis assumes that all markets clear.

The economy is then at full employment and potential output. A monetary expansion will raise prices but not output, and a fiscal expansion will crowd out private consumption and investment until aggregate demand is restored to potential output. At the other extreme, Keynesian analysis assumes that markets, especially the labour market, do not clear. With imperfect wage flexibility, a fall in aggregate demand for goods and the demand for labour reduce output and employment. Expansionary fiscal and monetary policy can increase real output.

Do markets clear or not? Before Keynes's *General Theory*, most economists assumed that markets cleared and tried to explain periods of high unemployment within this framework. In the immediate post-war period, most economists assumed that markets did not clear continuously and interpreted macroeconomics within the Keynesian paradigm.

After 1970, the pendulum swung back again. Many economists argued that, if wage stickiness leads to involuntary unemployment, workers will find a way to make wages more flexible, avoiding involuntary unemployment. People said the Keynesian assumption of wage stickiness could not be given plausible microeconomic foundations. In the last two decades, the pendulum has been in motion again. New Keynesian economists have developed micro-foundations for wage stickiness, and fewer economists presume markets automatically clear.

The attempt by some economists to explain even short-run fluctuations with market-clearing models has spawned a new literature on what determines potential output and equilibrium unemployment, topics neglected when the focus of analysis was simply movements in aggregate

demand. It is now generally accepted that movements in potential output may be significant, even in the short run. Whether they are the *only* source of short-run output fluctuations is essentially the same question as whether market clearing can be assumed, even in the short run.

Is long-run equilibrium unique?

Suppose an economy in long-run equilibrium then experiences a *temporary* shock which drives it to a different position in the short run. What happens once the shock disappears? Does the economy, sooner or later, go back to the original equilibrium or settle down in a new, *permanently different*, long-run equilibrium?

The latter case is *hysteresis*. Hysteresis exists when the path that an economy follows in the short run affects which long-run equilibrium it eventually reaches. There are then several possible equilibria in the long run. Chapter 23 discussed mechanisms that may give rise to hysteresis.

Whether hysteresis is quantitatively important is a controversial issue. If hysteresis matters, the easiest way to prevent its damaging effects is to prevent the economy from entering a recession in the first place. In contrast, economists who believe that hysteresis is unimportant take a more relaxed attitude to temporary recessions, which have no long-term consequences.

Expectations formation

Most economists accept that beliefs about the future affect behaviour today. Such beliefs certainly affect consumption and investment demand. Some disagreements between economists can be traced to different beliefs about how expectations are formed. For simplicity, we divide the possible approaches into three categories.

Exogenous expectations are determined outside the model and are an input to the analysis.

Exogenous expectations Some economists are agnostic on the vital question of how expectations are formed. They are treated as exogenous, or given, inputs to the analysis. The analysis shows *consequences* of a change in expectations—for example, a rise in expected future profits raises investment spending at each interest rate—but the analysis does not

investigate the *cause* of the change in expectations. In particular, it is unrelated to other parts of the analysis. With given expectations, there is no automatic feedback from rising output to expectations of higher profits in the future.

At best, exogenous expectations give an incomplete account of how the economy works. At worst, they neglect some inevitable feedback from the variables they are analysing to the expectations that were an input to the analysis.

Extrapolative expectations A simple way to make expectations endogenous, or determined by what is going on elsewhere in the analysis, is to assume people forecast future profits by extrapolating the behaviour of profits in the recent past, or extrapolate past inflation in order to form expectations of inflation in the near future. Proponents of this approach suggest that it offers a simple rule of thumb and corresponds to what many people seem to do in the real world.

Rational expectations Suppose half the world oil supply is destroyed by a war. You could use simple economics (supply and demand) to guess that oil prices will jump up sharply. You should raise your forecast of oil prices immediately. If you merely extrapolate past growth of oil prices, you will keep mistakenly under-forecasting future oil prices. It is implausible to keep using a forecasting rule that makes the same mistake period after period.

Extrapolative expectations assume that the future is an extension of the recent past.

Rational expectations guess the future correctly on average.

Forecasting rules that systematically give too low a forecast or too high a forecast are not used. Any tendency for expectations to be systematically in error is quickly detected and put right. We live in a risky world where unforeseeable things are always happening. Expectations are fulfilled only rarely. Rational expectations make good use of information available today, and do not make forecasts that are already knowably wrong. Only genuinely unforeseeable things make present forecasts go wrong. Sometimes people under-predict; sometimes they over-predict. Any systematic tendency to do one or other gets noticed and the basis of expectations formation is amended until guesses are on average correct.

Short run and long run

Where policies have short-run benefits but long-run costs, or vice versa, different groups of economists may adopt differing value judgements about how these gains and losses should be traded off. In part, the differing policy prescriptions offered by different economists reflect differing judgements about the relative importance of the short run and the long run.

The more quickly one believes markets clear, the less scope there is for demand management in the short run and the greater the importance of supply-side policy to raise potential output in the long run. Conversely, the more one believes in the possibility of high levels of Keynesian unemployment in the short run, the more likely one is to judge that the short-run benefits of returning to full employment outweigh any tendency thus induced to reduce potential output in the long run. Similarly, the more one's horizon is short-run, the more plausible it becomes that expectations can be treated as exogenous. The more one wants to discuss the long run, the more important it is to model how expectations are changing over time. The more one believes in hysteresis, the more one must look after the short run to look after the long run.

Contemporary macroeconomic thought

Having identified four areas of disagreement, we now examine the major schools of contemporary macroeconomic thought.

New Classical macroeconomics

The analysis is *classical* because it assumes that wage and price flexibility restore the economy to its position of full employment and potential output. The analysis is *new* because it assumes that expectations adjustment, as well as wage and price adjustment, is almost instant. At best, monetary and fiscal policy affect the *composition* of full-employment aggregate demand. Its *level* is necessarily potential output. This being unique, hysteresis is unimportant.

New Classical macroeconomics is based on the twin principles of rapid market clearing and rational expectations.

Wage and price adjustment is almost instant. Whatever level of unemployment is observed is thus the natural rate of unemployment. Unemployment changes over time because microeconomic incentives alter the natural rate itself.

Because expectations are rational, the government cannot use fiscal and monetary policy systematically to fool people. Suppose the government switches to a more expansionary monetary policy. This tends to make prices rise, since the economy begins close to full employment. If the initial policy change was not foreseen, workers will not have foreseen the price rise. They have settled for too low a money wage. Firms temporarily have cheap labour and expand output. Unanticipated monetary expansion causes an unanticipated rise in output and employment, above their long-run levels.

But if everyone has rational expectations, people quickly catch on to what the government is doing. When wages are renegotiated, everyone knows the money supply is expanding and prices are rising. The next nominal wage settlement suitably reflects this and, in the absence of any further surprises, real wages are at their equilibrium level again.

It is only the fact that some variables, particularly nominal wages, must be set in advance that prevents continuous full employment and potential output. Variables set in advance are set at levels expected to produce full employment. Only unexpected developments make them temporarily inappropriate, allowing output and employment to depart temporarily from their natural rates. But the government cannot use fiscal and monetary policy to make prices unexpectedly high period after period, and thus cannot hold output systematically above potential output. Essentially, demand management through monetary and fiscal policy is completely impotent.

It only remains for the government to control the price level and to pursue supply-side policies to raise potential output. Supply-side policies include income tax cuts to increase the incentive to work. Tight monetary policy will keep inflation under control. Low government spending will prevent large government borrowing from bidding up interest rates and crowding out private investment.

Nor will tight fiscal and monetary policy cause Keynesian unemployment. Wages and prices adjust to restore aggregate demand to potential output. If a switch to tighter policy takes people by surprise, at worst it has only temporary effects on output and unemployment. As soon as wages are renegotiated, they adjust to restore full employment. Thus New Classical

economists believe not merely that long-term trends but also short-run fluctuations have little to do with aggregate demand.

Real business cycle theorists belong to the same family as the New Classical macroeconomics, although their emphasis is a little different. Both believe in near-continuous market clearing and rational expectations. The New Classicals stress the effects of temporary surprises until expectations quickly catch up, thus developing a theory of fluctuations around potential output. Real business theorists take this a stage further and seek to explain all fluctuations as fluctuations in potential output itself.

Thus, the real business cycle approach is both more extreme and more general than that of New Classical macroeconomics. It is more extreme because its analysis neglects deviations from potential output even for a short time. Since changes in nominal money have no real effects in such a context, the cause of changes must be sought in shocks to real variables such as technical knowledge.

The approach is more general than that of New Classical macroeconomics because it concentrates all its powers of analysis on making explicit the microeconomic foundations for the intertemporal decisions of firms, households and governments. It is in decisions to amend intertemporal plans and reallocate them over time that real business cycle theorists believe they can explain how large movements in actual output and employment could be movements in equilibrium output and employment.

Gradualist monetarists

This school is associated with the Chicago professor, Milton Friedman (1912–2006). We use the term ‘monetarist’ to mean those economists espousing the classical doctrine that an increase in the money supply leads essentially to an increase in prices rather than to an increase in output. Thus, the New Classical economists believe in almost instant monetarism.

Gradualist monetarists believe that full employment is restored within a few years, so the main effect of higher money is higher prices.

New Classical economists believe in only temporary departures from full employment as a result of unforeseeable shocks that cannot immediately be reflected in wages. **Gradualist monetarists** accept that restoration of full employment takes a little longer. Even so, they believe that within a few

years wage and price adjustment *will* restore full employment. Like the New Classical economists, Gradualist monetarists do not believe that hysteresis matters. When the economy gets back to full employment after a temporary shock, it returns to the *same* long-run equilibrium in real terms.

The school believes in some of the arguments for wage rigidity presented in Chapter 23, but only for a short time. Different members of this school adopt different assumptions about expectations formation. Sluggish adjustment in expectations formation may provide an extra reason for slower adjustment back to full employment.

Gradualist monetarists believe that, in the short run, a fiscal or monetary stimulus *would* alter aggregate demand, output and employment, but that it is neither sensible nor desirable to undertake such policies. The short run must be subordinated to the interests of the long run.

Since wage and price adjustment take a few years to complete, expansionary monetary or fiscal policy can increase aggregate demand, output and employment in the short run. However, the Gradualists offer two reasons why policy should not be used in this way.

First, the economy will automatically return to full employment within a few years anyway. In the long run, trying to keep output above potential output leads only to inflation. Second, if instead the aim of policy is to react to shocks and reduce fluctuations around potential output, the policy may be counterproductive. By the time a shock is diagnosed and the necessary action taken, the economy may already be expanding on its own as wage and price adjustments begin to lead it back to full employment. Stabilization policy may exacerbate cycles not dampen them.

Since departures from full employment last a relatively short time, it is on the long-run classical analysis that the Gradualists place the most emphasis. The government's chief responsibility is to raise potential output through supply-side policies and the pursuit of price stability.

Moderate Keynesians

Broadly speaking, this group are short-run Keynesians and long-run monetarists.

In the short run, a fall in aggregate demand can generate a significant recession. Although many economists in this group believe that expectations adjustment is also sluggish, some of them believe in rational

expectations and hold that it is not systematic mistakes in expectations formation but sluggish wage and price adjustment that prevent rapid restoration of full employment. By sluggish we mean that they do not respond quickly to departures from potential output and equilibrium unemployment. Nominal wages may still change rapidly because of expected inflation.

Moderate Keynesians believe the economy will eventually return to full employment, but that this could take many years.

Moderate Keynesians believe that recessions last a bit longer than the couple of years over which a Gradualist monetarist believes markets unaided can restore full employment. Hence, Moderate Keynesians draw a different judgement about the relative importance of the short run and the long run. Slower market adjustment reduces the danger that, by the time government has diagnosed the problem, the market is already fixing it. Slower adjustment also raises the need for stabilization policy. Thus Moderate Keynesians believe that the government should accept responsibility for stabilization policy in the short run.

Since Moderate Keynesians believe the economy will *eventually* return to full employment, they accept that persistent rapid monetary growth must eventually lead to inflation once the full employment position is reached. In the very long run, only supply-side policies can generate sustained economic growth by raising potential output. Thus many economists in this group argue that the government should not neglect two of the policy prescriptions of the monetarists. Supply-side policies are important in the long run; and, if high inflation reduces potential output, in the long run the average level of fiscal and monetary policy must be compatible with low inflation.

Moderate Keynesians see no conflict between this stance of policy in the long run and the recommendation that in the short run active stabilization policies should be undertaken. Credible policy makers can be active precisely because people trust that their actions will be temporary not permanent. If a current stimulus is reversed as soon as the crisis is over, it need not threaten price stability in the medium run.

New Keynesians

As explained in Concept 27.1 above, New Keynesians have tried to provide more rigorous micro-foundations for Keynesian analysis while espousing many of the tools (rational expectations, stochastic dynamic general equilibrium) that had been adopted by the New Classical and real business cycle attack on more primitive Keynesianism. As such, they belong in the Moderate Keynesian camp. They recognize the necessity of keeping track of aggregate demand as well as aggregate supply. But they continue to believe that many shocks have their origins (and solutions) in shifts in aggregate demand.

Extreme Keynesians

Extreme Keynesians believe markets do not clear, even in the long run.

Keynesian unemployment may persist indefinitely unless the government intervenes to boost aggregate demand. Extreme Keynesians reject the view that slumps can eventually restore full employment via downward pressure on wage growth and inflation.

Whereas hysteresis suggests that, once a recession has done its damage, supply has then been eliminated, so boosting demand no longer works, Extreme Keynesians believe that boosting aggregate demand through government policy will do the trick.

This case rests primarily on labour market rigidity. Real-wage rigidity causes excess supply in the labour market; that is, a pool of involuntarily unemployed workers that remains available at any time to be mopped up through demand expansion. Extreme Keynesians refer to this assumption of labour market inflexibility as the *real-wage hypothesis*.

Why can't all nominal variables fall, reducing inflation and allowing the central bank to reduce real interest rates, thereby eventually moving the economy to full employment? Extreme Keynesians have several answers.

First, it is impossible to co-ordinate the fall in wage growth and inflation. If all rates of wage growth could be cut together, no real wage need change. But in practice, some workers have to go first. Unless and until all other wage and price growth slows down, the first workers to reduce the growth of nominal wages also cut real wages. This may be sufficient to prevent the wage cut taking place, especially if each group of workers is very sensitive about its wages relative to other groups.

Second, the central bank can cut the nominal interest rate to zero but no further. Hence, if a recession is deep enough to induce negative inflation, it then raises real interest rates, exacerbating the recession further. Third, when times are tough and firms are losing money, they do not wish to invest, even at zero interest rates. Old Keynesians used to compare monetary policy to a string: you can pull tight on it in a boom, but pushing on it in a slump may have no effect. Thus, Extreme Keynesians stress the role of fiscal policy in getting the economy out of a serious recession.

Just as New Classical economists are optimists about both the speed of market clearing and the ability of people intelligently to form, and rapidly to adjust, expectations about the future, Extreme Keynesians are nearly as pessimistic about expectations as they are about market clearing. Keynes himself compared expectations to a beauty contest. The modern equivalent would be a TV game show where the competitor has to guess the answer most frequently chosen by the TV audience.

In such situations, what matters is not getting the right answer (which is how economists try to evaluate rational expectations): what matters is guessing what other people guess. Multiple equilibria may be common, which undermines the ease with which we can assume that people quickly adjust expectations to *the* right answer. Through Extreme Keynesian spectacles, co-ordination failures (externalities) occur as much in expectations as in wage-setting.

Summing up

We have set out the views of the competing schools of modern macroeconomics. In each case, we have sought to interpret their views against four basic assumptions: about market clearing, about expectations formation, about hysteresis and about the relative priority given to short run and long run. Table 27.2 summarizes our discussion.

We did not adjudicate between the competing views of macroeconomics, though we are probably in the Moderate Keynesian group of economists. Rather, we sought to develop a framework in which the differing positions can be interpreted. We have explained how changes in the basic assumptions, especially about the speed of adjustment, the time required for restoration of full employment and the possibility of hysteresis, allow this framework to reflect the views of the different schools of modern macroeconomics, and show why they reach differing policy recommendations.

Table 27.2 Schools of macroeconomic thought

Issue	New Classical	Gradualist Monetarist	Moderate Keynesian	Extreme Keynesian
Market clearing	Very fast	Quite fast	Quite slow	Very slow
Expectations Adjustment	Rapid	Slower	Fast or slow	Slow
Long run/short run	Little difference since adjust fast	Long run more important	Don't forget short run	Short run vital
Full employment	Always close	Never far away	Could be far away	Could stay away
Hysteresis	No problem	No problem	Might be problem	Problem
Demand management or supply side policy	Forget demand; supply side needed	Supply more important; avoid swings in demand	Demand matters too	Demand what counts

Summary

- The **trend path of output** is the long-run path after short-run fluctuations are ironed out. The business cycle describes fluctuations in output around this trend. Cycles last about five years but are not perfectly regular.
- A **political business cycle** arises from government manipulation of the economy to make things look good just before an election.
- **Persistence** requires either sluggish adjustment or intertemporal substitution. Persistence is necessary but not sufficient for cycles.
- The **multiplier-accelerator model** assumes investment depends on expected future profits, which reflect past output growth. The model delivers a cycle but assumes that firms are stupid: their expectations neglect the cycle implied by their own behaviour.
- Full capacity and the impossibility of negative gross investment provide **ceilings and floors** that limit the extent to which output can fluctuate.

- Fluctuations in **stockbuilding** are important in the business cycle. The need to restore stocks to original levels explains why output continues to differ from demand even during the recovery phase.
- **Real business cycles** are cycles in potential output itself. In such circumstances, it is not desirable for policy to dampen cycles.
- Some swings in potential output do occur, but many short-run fluctuations probably reflect Keynesian departures from potential output. Aggregate demand and aggregate supply both contribute to the business cycle.
- Increasing integration of world financial and product markets has made most countries heavily dependent on the wider world. Business cycles in the rich countries are closely correlated.
- There is much about which all economists agree. There are also differences of opinion, both in the positive economics of how the world actually works and in the normative economics of how the government should behave.
- Economic theories should be tested against the facts. In some cases, tests do not yield conclusive answers. Some variables, such as expectations, are unobservable. The world is also changing. It may be impossible to get enough data on the world as it is today to allow definitive empirical tests of competing theories.
- The major **schools of macroeconomic** thought can be viewed in relation to four key issues: the speed with which the labour market clears, how expectations are formed, the possibility of hysteresis and the relative importance of the short run and long run.
- **New Classical macroeconomists** assume market clearing is almost instant. Only predetermined contracts prevent continuous full employment. **Rational expectations** embody the best guess at the time about future values. Any foreseeable change is already built into these variables. Only pure surprises cause temporary departures from full employment until preset variables can be altered and full employment

restored. With the economy near potential output, demand management is pointless. Government policy should minimize surprises. Surprises apart, movements in output reflect movements in potential output. Policy should pursue price stability and supply-side policies to raise potential output.

- **Real business cycle theorists** neglect even temporary departures from full market clearing. They argue that intertemporal decisions of households, firms and government can explain even short-term fluctuations as movements in potential output.
- **Gradualist monetarists** believe that restoration of potential output, though not instant, takes only a few years. Attempts at demand management may be counterproductive if the economy is already recovering by the time a recession is diagnosed. The government should not ‘fine-tune’ aggregate demand but concentrate on long-run policies to keep inflation down and promote supply-side policies to raise potential output.
- **Moderate Keynesians** believe automatic restoration of full employment can take many years but will happen eventually. Although demand management cannot raise output without limit, active stabilization policy is worth undertaking to prevent booms and slumps that could last several years and therefore are diagnosed relatively easily. In the long run, supply-side policies are still important, but eliminating big slumps is important if hysteresis has permanent effects on long-run equilibrium.
- **New Keynesians** provide microeconomic foundations for Keynesian macroeconomics, based principally on costs of changing prices and wages. Several channels for hysteresis have also been developed.
- **Extreme Keynesians** believe departures from full employment may be protracted. Keynesian unemployment does not make real wages fall, and may not even reduce inflation. Even if it does, aggregate demand may not respond to lower interest rates if pessimism is high. The first responsibility of government is not supply-side policies to raise potential

output that is not attained anyway, but restoration output to potential output by expansionary fiscal and monetary policy, especially the former.

Review questions



EASY

- 1 Would it help the world economy if all the largest countries elected governments on the same day? Why, or why not?
- 2 'If firms could forecast future output and profits accurately, there would not be a business cycle.' Is this true?
- 3 Heavily dependent on output of oil and fishing, Norway's business cycle goes the other way from that in other European countries. Why?
- 4 **Common fallacies** Why are these statements wrong? (a) Closer integration of national economies will abolish business cycles. (b) The more we expect cycles, the more we get them. (c) Because output and labour productivity are closely correlated, fluctuations in productivity are the main cause of business cycles.

MEDIUM

- 5 Which of the following statements are correct? (a) Business cycles imply people do not expect fluctuations in the economy: if they could see a cycle coming, they would already be taking action to abolish it. (b) It is easy to explain why the economy's return to long-run equilibrium takes time, but it is not possible to explain why this return causes actual output to overshoot potential output. (c) Economic dynamics are slow and complicated. There are many models explaining the dynamics in the economy which cause business cycles.
- 6 Why might voters care more about the direction in which the economy is heading than about the absolute level of its position at election time?
- 7 (a) Since central banks became independent, do you expect to see more or less evidence of a political business cycle? (b) Might there be an interest rate cycle instead? Why, or why not?
- 8 Suppose $Y = C + I$, $C = A + 0.6Y$ and $I = 0.1(\Delta Y)$. Does this economy converge to long-run equilibrium, explode away from long-run equilibrium or cycle forever?
- 9 If the multiplier-accelerator model still fits the data quite well, does this imply that people are stupid?

|0 Greece, Spain, Portugal, Ireland and Italy have emerged as weak members of the Eurozone. Do you think this is because their business cycles are less correlated with France and Germany or because their political institutions are weak?

|1 What do real business cycles explain?

HARD

|2 Consider an economy with a fixed exchange rate. Beginning from internal and external balance, the economy experiences an adverse domestic demand shock that is not fully offset by a policy response. Draw a diagram to illustrate subsequent adjustment. Why does the requirement to get back to *both* internal and external balance generate a cyclical response?

|3 Plot the data in the table below and confirm that both output and investment exhibit cyclical behaviour. Which is causing which?

The multiplier-accelerator model of the business cycle

Period	Change in last period's output ($\gamma_{t-1} - \gamma_{t-2}$)	Investment I_t	Output γ_t
$t = 1$	0	10	100
$t = 2$	0	10	120
$t = 3$	20	20	140
$t = 4$	20	20	140
$t = 5$	0	10	120
$t = 6$	-20	0	100
$t = 7$	-20	0	100
$t = 8$	0	10	120
$t = 9$	20	20	140

|4 Essay question ‘The business cycle ought to last for different lengths of time depending on whether the original shocks were supply shocks or demand shocks.’ Is this true?

1 For an accessible introduction to these issues, see the lively exchange between Charles Plosser and Greg Mankiw, ‘Real business cycles: A new Keynesian perspective, *Journal of Economic Perspectives* 3, no. 3 (1989): 79–90.

2 ‘Impact of the current economic and financial crisis on potential output, *European Economy*, Occasional Paper 49, June 2009.

CHAPTER 28

Supply-side economics and economic growth

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 explain supply-side economics
- 2 understand growth in potential output
- 3 describe Malthus forecast of eventual starvation
- 4 understand how technical progress and capital accumulation made the forecast wrong
- 5 describe the neoclassical model of economic growth
- 6 explain the convergence hypothesis
- 7 analyse the growth performance of rich and poor countries
- 8 understand whether policy can affect growth
- 9 understand whether growth must stop to save the environment

Much of Part Four has focused on the causes and consequences of changes in the demand for output and labour. We first introduced aggregate supply in Chapter 21. Sluggish adjustment of wages and prices induces a distinction between short-run supply and long-run supply. Chapter 21 and Chapter 27 discussed how the economy makes the transition from short run to long run. Supply-side economics analyses how to increase aggregate supply through better incentives or greater efficiency.

This chapter is entirely about aggregate supply in the long run, and hence about the paths of potential output and equilibrium unemployment. We

begin by analysing what actions can be taken to achieve a one-off increase in aggregate supply.

Most supply-side policies are microeconomic rather than macroeconomic. We have already referred to several, for example in Chapter 23 when discussing equilibrium unemployment. That chapter also introduced the concept of hysteresis, in which temporary macroeconomic phenomena have lasting supply-side effects. This is one case in which macroeconomics matters for aggregate supply. In this chapter, we pull together different strands to provide a comprehensive analysis of supply-side economics.

Supply-side economics analyses how to increase aggregate supply through better incentives or greater efficiency.

The longer the period that we analyse, the less important one-off changes become. In the very long run, *continuing* increases in output are much more important. The second part of the chapter analyses causes of *economic growth*, inducing increases in potential output year after year. Essentially, this requires the steady accumulation of something lasting – human capital, physical capital or technical knowledge – that provides each generation with a better inheritance than its predecessors.

28.1 Supply-side economics

We begin with one-off changes. It is useful to distinguish between those aiming to increase labour input, and those aiming to increase output per unit of labour input.

Higher labour input

Broadly, higher labour input falls into four categories: higher labour force participation, a higher fraction of the labour force accepting a job, longer hours worked on the job and more effort by workers while working. None can increase without limit – they are all sources of one-off effects.

In most advanced economies, labour force participation rates are high for men and have risen dramatically for women in the last 50 years. Governments would usually like participation to be even higher, both to

boost potential output and to reduce the welfare bill that supports the old, the sick and those not bothering to look for work. However, it may also be important for young children to spend time with at least one parent. There are limits to efforts to raise labour force participation. Similar arguments apply to attempts to increase labour supply by raising the job acceptance rate or the number of hours worked.

Section 10.4 emphasized the need to consider both income and substitution effects when discussing how incentives affect both decisions to join the labour force and how many hours to work if a job offer is accepted. Higher take-home pay, relative to welfare benefits, makes working more attractive (the substitution effect) but also increases the demand for leisure (making working less attractive).

The wide range of other possible means-tested welfare benefits, from child support to housing subsidies, greatly complicates the analysis. Every government thinks it has found the secret of making work pay. In practice, they are often disappointed with the results of their policy interventions. It is difficult to be brutal and compassionate at the same time. Section 23.3 discusses in more detail the effects of welfare benefits and tax rates on decisions to join the labour force and to accept a job.

The interplay of substitution and income effects will often lead to disappointing small effects of attempts to increase incentives to work longer; the same conclusion applies to attempts to induce workers to contribute greater effort while at work.

Chapter 23 also introduced hysteresis. A temporary recession may cause capital to be scrapped, activities to be abandoned and workers to leave the labour force. Even once demand is restored, supply may be permanently reduced. Section 27.4 discussed recent estimates of the lasting adverse supply-side consequences of the financial crash and subsequent recession.

Increasing labour productivity

A second channel by which potential output can be increased is by increasing output per worker. Two important channels for increasing labour productivity are through technical progress and by providing more physical capital for each worker to use. These can be one-off or ongoing. The latter forms the basis of the long-run theory of economic growth, discussed later in the chapter.

The quantity of physical capital available depends in part on the willingness of the country to save rather than consume. The technology available depends on scientific discovery (both in universities and in research labs funded by companies or the government), converting invention to adoption through the practical process of innovation, having intellectual property laws that protect rewards for innovators, and the effectiveness of take-up of new ideas. Policies to stimulate saving and to promote research and development (R&D) can thus be viewed as supply-side policies.

However, there are many other ways in which productivity gains can be achieved. First, the human capital of each worker can be enhanced through investment in education and training. This should not be narrowly interpreted as the skills acquired at work, but much more broadly as including literacy, numeracy, IT familiarity and having acquired a work ethic. Societies that bring up young workers to have these attributes have higher productivity than those which do not.

Work and organizational practices may also have significant effects. When Japanese car companies built factories in the UK, their productivity was significantly higher than that in UK-owned factories. This was not simply because the Japanese invested in better capital for automating car production. They also organized factories differently: workers and management ate together, quality circles allowed workers' ideas for improvement to feed quickly into production improvements and the newly hired UK workforce had incentives aligned with those of the company. There were no vested interests to oppose change in order to defend people's own jobs. Social, organizational and cultural capital can sometimes be as important as physical capital.

More generally, even when it is known that a more efficient production solution exists, some particular group may lose out and have an incentive to oppose that change. Some societies are better than others at coping with change. Generally, those that change more quickly are at one end of the spectrum – either largely free market, in which case the losers from change cannot successfully oppose it, or largely centralized, in which case the powerful centre can force through change despite opposition from potential losers. The United States is close to the former, China to the latter.

Societies in the middle – such as the social democratic models of Western Europe – often create governments insufficiently powerful to

force through change, while having sufficiently powerful market participants (trade unions, cartels) that can block changes not in their own interest. Europe's high level of productivity is well explained by its high levels of education and inherited capital stock. For several decades, it has not been particularly successful at increasing productivity.

Failing to allow unsuccessful companies to die inhibits 'natural selection'—in which the strong do better and the weak are weeded out. Enhancing competition may have a similar effect. Yet unfettered competition and unregulated markets are not always the answer. Just think of the damage done by lightly regulated bankers, largely operating in their own self-interest, whose actions led directly to the financial crash with huge and ongoing costs for the world economy.

The right amount of regulation, effectively enforced, is desirable but hard to achieve in practice. Governments, afraid of the political consequences of high tax rates, often underfund regulatory agencies. In contrast, large companies have every incentive to invest in learning how to circumvent existing regulations.

Beginning from a heavily regulated economy, deregulation may achieve greater competition and efficient gains from natural selection. Beyond some point, further deregulation adversely affects the supply side by leading to opportunistic behaviour that is not in the general interest. Knowing exactly where to draw the line is difficult.

Summing up

Everyone wants to improve the supply side. In reality, supply-side improvement is difficult. Even a one-off increase in potential output, if truly permanent, is a considerable achievement. All of the examples above have been tried, sometimes successfully and sometimes with disappointing results. Supply-side policies are often controversial, since they usually have sharp implications for redistribution. Cutting welfare benefits may encourage incentives to work, but some people think that there are limits to how far a civilized society should proceed in this direction. Clamping down on the behaviour of private firms may prevent excesses but also risks stifling the innovation that drives productivity growth.

We now turn from the analysis of one-off changes in potential output to the analysis of continuing growth in potential output.

28.2 Economic growth: preliminary remarks

Figure 28.1 shows data on 13 OECD countries. In 1870, per capita real GDP (in 2011 US dollars) ranged from \$500–3000. By 2011, it ranged from \$18 000 to over \$30 000 per person. On average, we are richer than our grandparents, but less rich than our grandchildren will be. Figure 28.1 prompts three questions. What is long-run economic growth? What causes it? And can economic policies affect it? We focus mainly on industrial countries that have grown a lot already.

Economists are fascinated by the theory of economic growth. In 1798 Thomas Malthus' *First Essay on Population* focused on the consequence of diminishing marginal productivity. As more and more labour became available, output would increase more slowly than employment, reducing output per person, thus causing starvation. An end to population growth and output growth would be the eventual result – the origin of economics as the ‘dismal science’. Some countries are still stuck in a Malthusian trap; others broke through to sustained growth and prosperity. We examine how they did so.

As Figure 28.1 implies, an extra 0.5 per cent on the annual growth rate makes a vast difference to potential output after a few decades. By the end of the 1960s, economists had worked out a theory of economic growth. It yielded many insights but had one central failing. It predicted that government policy made no difference to the long-run growth rate.

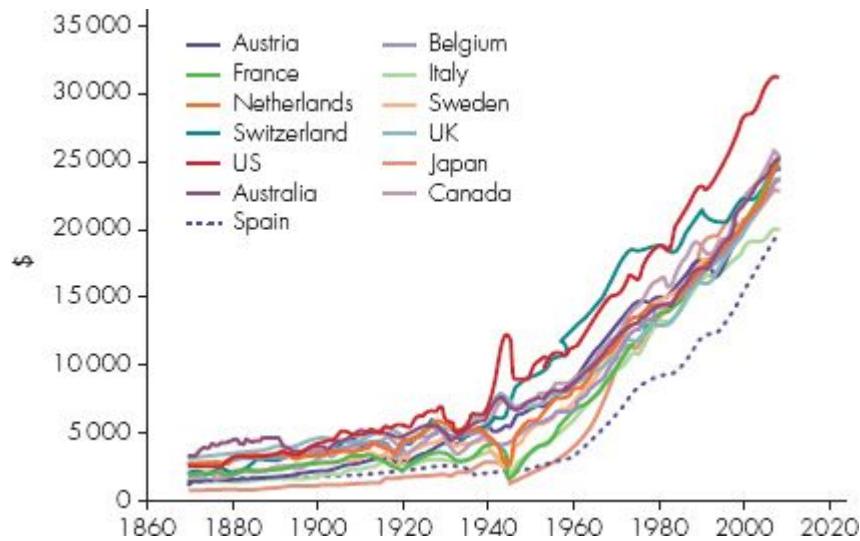


Figure 28.1 Per capita real GDP growth since 1870 (in constant 2011 US\$ prices)

Source: I. Kitov and O. Kitov, 'Real GDP per capita since 1970' (<http://ssrn.com/abstract=52066579>).

In the mid-1980s, a simple insight spawned a new approach in which long-run growth is affected by private behaviour and government policy. We briefly explain this new approach to **economic growth**.

Economic growth is the rate of change of real income or real output.

Finally, we consider whether growth is good. Might it be better to grow more slowly? Can the costs of growth outweigh its benefits?

The growth rate of a variable is its percentage rise per annum. To define economic growth, we must specify both the variable to measure and the period over which to measure it. Figure 28.1 used real GDP per head. We get similar results using per capita real GNP.

GDP and GNP measure the total output and total income of an economy. But they are very incomplete measures of *economic* output and income.

GDP and economic output

GDP measures the net output or value added in an economy by measuring goods and services bought with money. It omits output not

bought and sold and therefore unmeasured. Two big omissions are leisure and externalities such as pollution or congestion.

In most industrial countries, average hours of work have fallen at least ten hours a week since 1900. In choosing to work fewer hours, people reveal that the extra leisure is worth at least as much as the extra goods that could have been bought by working longer. When people swap washing machines for extra leisure, recorded GDP falls. GDP understates the true economic output of the economy. Conversely, the output of pollution reduces the net economic welfare that the economy is producing, and ideally should be subtracted from GDP.

Including leisure in GDP would have raised recorded GDP in both 1870 and 2011. Since the value of leisure probably rose less quickly than measured output, which rose 11-fold in the UK and 100-fold in Japan, a more comprehensive output measure might show a slower growth rate.

Conversely, pollution and congestion have increased rapidly. Allowing for them would also reduce true growth rates below those shown in Figure 28.1. A measure of true economic output each year would have to allow for environmental depreciation – everything from the true cost of global warming to the reduction in genetic diversity and the loss of amenities as grasslands are replaced with urban sprawl.

New products

In 1870 people had no TVs, cars or mobile phones. Statisticians do their best to compare the value of real GDP in different years, but new products make it hard to compare across time. We can estimate how much people's real income rises when a new product does an old task more cheaply. The calculation is harder when the new product allows a new activity not previously possible. A small amount of what we think of as inflation probably reflects real price increases justified by better quality or completely new products.

GDP and happiness

Even with an accurate and comprehensive measure of GDP, two problems remain. First, do we care about total GDP or GDP per capita? This depends on the question we wish to ask. Total GDP shows the size of an economy. However, if we care about the welfare of a typical

individual in an economy, it is better to look at GDP per capita. Real GDP grew more quickly in Australia than in France or Sweden during 1870–2011; however, in part this reflected rapid population growth, largely through immigration. Sweden and France had faster growth in GDP per person over the period.

Real GDP per person is an imperfect indicator of the happiness of a typical citizen. When income is shared equally between citizens, a country's per capita real GDP tells us what every person gets. But some countries have very unequal income distributions. A few people earn a lot, and a lot of people earn only a little. Such countries may have fairly high per capita real income but many citizens still live in poverty.

Even when GDP is adjusted to measure leisure, pollution and so on, higher per capita GDP need not lead to greater happiness. Material goods are not everything. But they help. Movements in which people return to ‘the simple life’ have not had much success. Most of the poorer countries are trying to increase their GDP as quickly as possible.

A recent phenomenon

Figure 28.1 has one more implication. An annual growth rate of only 1.3 per cent in per capita GDP led to a 5.5-fold rise in UK per capita real GDP between 1870 and 2011. In 1870 UK per capita income was about £1900, measured in sterling and using 2000 prices. If its annual growth rate had always been 1.3 per cent, per capita real income would have been £370 in 1750, £75 in 1630 and £16 in 1510. This is implausible. It is only in the last 250 years that per capita real income has risen steadily.

In the long run, output fluctuations around potential output are swamped by the growth of potential output itself. If potential output rises 2 per cent a year, it will increase seven-fold in less than a century. To explain growth, we must think about changes in potential output.

28.3 Growth: an overview

The **production function** shows the maximum output obtainable from specified quantities of inputs, given the existing technical knowledge.

For simplicity, we assume that the economy is always at potential output. The [production function](#) tells us that higher potential output can be traced to more inputs of land, labour, capital and raw materials, or to technical advances that let given inputs make more output.

In the long run, population growth may be affected by per capita output, which affects the number of children people decide to have, and the health care and nutrition people then get. Nevertheless, we simplify by assuming that the rate of population growth is independent of economic factors. Anything that raises output then also raises per capita output.

Capital

Productive capital is the stock of machinery, buildings and inventories which, with other inputs, combine to make output. For a given labour input, more capital raises output. However, capital depreciates over time. Some new investment is needed just to stop the existing capital stock from shrinking. And with a growing labour force, even more investment is needed if capital per worker is to be maintained. With yet faster investment, capital per worker rises over time, increasing the output each worker can produce. Higher capital per worker is a key means of raising output per worker and per capita income.

Labour

Over a few decades, employment can rise because labour force participation is increasing or because equilibrium unemployment is falling. However, over a longer period, these one-off changes account for less and less of the total change in employment and output. However, sustained population growth is a candidate for a cause of continuing expansion of employment.

Human capital

Human capital is the skill and knowledge embodied in the minds and hands of workers. Education, training and experience allow workers to make more output. For example, much of Germany's physical capital was devastated during the Second World War but the human capital of its labour force survived. Given these skills, Germany recovered rapidly

after 1945. Without its human capital, there would have been no post-war German economic miracle.

Human capital can be accumulated over time and across generations. The young can absorb quickly the lessons learned more painfully by their forebears, enhancing their productivity as workers. Human capital is a candidate for a source of ongoing growth.

Land

Land is especially important in an agricultural economy. If each worker has more land, agricultural output is higher. Land is less important in highly industrialized economies. Hong Kong and Singapore have grown rapidly despite overcrowding and a scarcity of land. Even so, more land would help.

Increases in the supply of land are pretty unimportant to growth. In theory, land is the input whose total supply to the economy is fixed. In practice, the distinction between land and capital is blurred. By applying more fertilizer per acre, the effective quantity of farming land can be increased. With investment in drainage or irrigation, marshes and deserts can be made productive. Dubai built superstar homes, hotels, and even a new airport, on land reclaimed from the sea. Increases in the supply of land are pretty unimportant to sustained growth.

Raw materials

Given the quantity of other inputs, more input of raw materials allows more output. When raw materials are scarce and expensive, workers take time and care not to waste them. With more plentiful raw materials, workers work more quickly. Depletable resources can be used only once.

When a barrel of oil has been extracted from the ground and used to fuel a machine, the world has one less barrel of oil reserves – it is a **depletable resource**. If the world has a finite stock of oil reserves, it will eventually run out of oil, though perhaps not for centuries.

Depletable resources can be used only once.

Renewable resources can be used again if not overexploited.

In contrast, timber and fish, if harvested in moderation, are replaced by nature and can be used as production inputs for ever – they are **renewable resources**. However, if over-harvested they become extinct. With only a few whales left, whales find it hard to find partners with whom to breed. The stock of whales falls.

Factor contributions and scale economies

The marginal product of a factor is extra output when that input rises by a unit but all other inputs are held constant. Microeconomics tells us that marginal products eventually decline as the input increases. With two workers already on each machine, another worker does little to raise output.

Instead of increasing an input in isolation, suppose all inputs are doubled together. If output exactly doubles, there are *constant returns to scale*; if output more (less) than doubles, there are *increasing (decreasing) returns to scale*.

Scale economies reinforce growth. Any rise in inputs gets an extra bonus in higher output. There may be engineering reasons for scale economies. Simple mathematics shows that it takes less than twice the steel input to build an oil tanker of twice the capacity. On the other hand, many developing countries regret that their resources are tied up in huge steel mills that are now inefficient. Bigger is not always better. In practice, economists often assume constant returns to scale.

Having discussed the different production inputs, we turn now to the role of technical knowledge.

28.4 Technical knowledge

At any given time, a society has a stock of technical knowledge about ways in which goods can be produced. Some of this knowledge is written down in books and blueprints, but much is reflected in working practices learned by hard experience.

Invention and innovation

Technical advances in productivity come through **invention**, the discovery of new knowledge, and **innovation**, the incorporation of new knowledge into production techniques.

Major **inventions** can lead to spectacular increases in technical knowledge. The wheel, the steam engine and the modern computer are examples. Technical progress in agriculture has also been dramatic. Industrial societies began only when productivity improvements in agriculture freed some of the workforce to produce industrial goods without leaving people short of food. Before then, everyone had to work the land merely to get enough food to survive. The replacement of animal power by machines, the development of fertilizer, drainage and irrigation, and new hybrid seeds, all played a large part in improving agricultural production and enabling economic growth.

To introduce new ideas to actual production, **innovation** often requires investment in new machines. Without investment, bullocks cannot be transformed into tractors even once the know-how for building tractors is available. Major new inventions thus lead to waves of investment and innovation as the ideas are put into practice. The mid-nineteenth century was the age of the train and the mid-twentieth century the age of the car. We are now in the age of the microchip.

Human capital can matter as much as physical capital. With practice, workers get better at doing a particular job. The most famous example is known as the Horndal effect, after a Swedish steelworks built during 1835–36 and kept in the same condition for the next 15 years. With no change in the plant or the size of the labour force, output per worker-hour nevertheless rose by 2 per cent a year. Eventually, however, as skills become mastered, further productivity increases are harder to attain.

CASE 28.1

GROWTH AND COMPETITION

For centuries, per capita income growth was tiny. Most people were close to starvation. Now we take growth for granted. After 1750, industrialization changed everything. Capital and knowledge, accumulated by one generation, were inherited and augmented by the next generation. Why 1750? Mathematical and scientific ideas

reached a critical mass, allowing an explosion of practical spin-offs. Yet many pioneers of the industrial revolution were commonsense artisans with little scientific training. Conversely, the ancient Greece of Pythagoras and Archimedes achieved scientific learning but not economic prosperity.

By the start of the fifteenth century, China understood hydraulic engineering, artificial fertilizers and veterinary medicine. It had blast furnaces in 200 BC, 1500 years before Europe. It had paper 1000 years before Europe, and invented printing 400 years before Gutenberg. Yet by 1600 China had been overtaken by Western Europe, and by 1800 had been left far behind.

Economic historians continue to debate the root causes of progress, but three ingredients seem crucial: values, politics and economic institutions. Growth entails a willingness to embrace change. China's rulers liked social order, stability and isolation from foreign ideas: fine attitudes when progress was slow and domestic but a disaster when the world experienced a profusion of new technologies and applications.

Powerful Chinese rulers could enforce bans and block change in their huge empire. When individual European rulers tried to do the same, competition between small European states undermined this sovereignty and offered opportunities for growth and change. Economic competition helped separate markets from political control. Rights of merchants led to laws of contract, patent, company law and property. Competition between forms of institution allowed more effective solutions to emerge and evolve. Arbitrary intervention by heads of state was reduced. Opportunities for business, trade, invention and innovation flourished.

The making of Western Europe

date	Per capita income (1990 prices)	Inventions
1000	400	Watermill
1100	430	Padded horse collar
1200	480	Windmill
1300	510	Compass
1400	600	Blast furnace

1500	660	Gutenberg printing press
1600	780	Telescope
1700	880	Pendulum clock, canals
1800	1280	Steam engine, spinning and weaving machines, cast iron, electric battery
1900	3400	Telegraph, telephone, electric light, wireless
2000	17 400	Steel, cars, planes, computers, nuclear energy

Source: adapted from The Economist, 31 December 1999. © The Economist Newspaper Limited 2010.

Research and development

What determines the amount of invention and innovation? Some new ideas are the product of intellectual curiosity or frustration ('There must be a better way to do this!'). But, like most activities, the output of new ideas depends to a large extent on the resources devoted to looking for them, which in turn depends on the cost of tying up resources in this way and the prospective benefits from success. Some research activities take place in university departments, usually funded at least in part by the government, but a lot of research is privately funded through the money firms devote to R&D.

The outcome of research is risky. Research workers never know whether or not they will find anything useful. Research is like a risky investment project. The funds are committed before the benefits (if any) start to accrue, but there is one important difference. Suppose you spend a lot of money developing a better mousetrap. When you succeed, everyone copies your new mousetrap: the price is bid down, and you never recoup your initial investment. In such a world, there would be little incentive to undertake R&D.

If the invention becomes widely available, society gets the benefit but the original developer does not: there is an *externality*. Private and social gains do not coincide and the price mechanism does not provide the correct incentives. Society tries to get round this *market failure* in two ways. First, it grants *patents* to private inventors and innovators – legal monopolies for a fixed period of time that allow successful research projects to repay investments in R&D by temporarily charging higher

prices than the cost of production alone. Second, the government subsidizes a good deal of basic research in universities, in its own laboratories and in private industry.

28.5 Growth and accumulation

In this section we explore the links between output growth, factor accumulation and technical progress. We organize our discussion around a simple production function:

$$Y = A \times f(K, L)$$

Variable inputs capital K and labour L combine to produce a given output $f(K, L)$. The function f tells us how much we get out of particular amounts of inputs K and L . This function f never changes. We capture technical progress separately through A , which measures the extent of technical knowledge at any date. As technical progress takes place, we get more output from given inputs: a rise in A . For simplicity, we assume that land is fixed.

Malthus, land and population

Writing in 1798 and living in a largely agricultural society, Malthus worried about the fixed supply of land. As a growing population worked a fixed supply of land, the marginal product of labour would fall. Agricultural output would grow less quickly than population. The per capita food supply would fall until starvation reduced the population to the level that could be fed from the given supply of agricultural land.

In terms of equation (1), starving people consume all their income. Without savings, society cannot invest in capital, so K is zero. The production function then has diminishing returns to labour: adding more workers drives down productivity. Figure 28.2 illustrates.

Some poor countries today face this *Malthusian trap*. Agricultural productivity is so low that everyone must work the land to produce food. As the population grows and agricultural output fails to keep pace, famine sets in and people die. If better fertilizers or irrigation improve agricultural output, the population quickly expands as nutrition improves, and people are driven back to starvation levels again.

Yet Malthus— prediction was not correct for all countries. Today's rich countries broke out of the Malthusian trap. How did they do it? First, they raised agricultural productivity (without an immediate population increase) so that some workers could be switched to industrial production. The capital goods then produced included better ploughs; machinery to pump water and drain fields; and transport to distribute food more effectively. As capital was applied in agriculture, output per worker rose further, releasing more workers to industry while maintaining enough food production to feed the growing population.

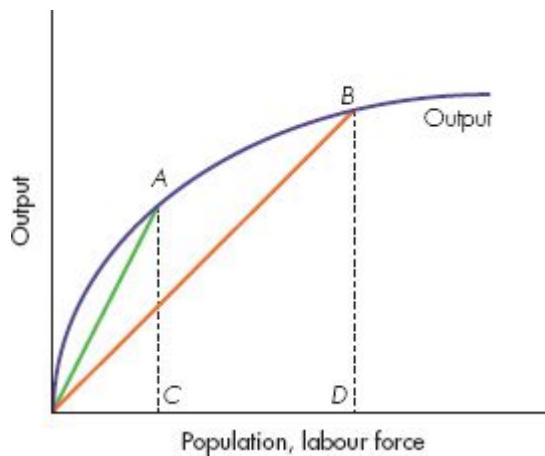


Figure 28.2 The Malthusian trap

The labour force grows with population, but there are diminishing returns to output. At labour force C, output is AC and output per head is given by the slope of OA. At population D, output is higher at DB but output per head has fallen from the slope OA to the slope OB. When output per head falls to starvation levels growth cannot continue.

Second, the rapid technical progress in agricultural production led to large and persistent productivity increases, reinforcing the effect of moving to more capital-intensive agricultural production. In terms of equation (1), rises in A and in K let output grow faster than labour, causing a *rise* in living standards.

Thus, even the existence of a factor in fixed supply need not make sustained growth impossible. If capital can be accumulated, more and more capital can be substituted for fixed land, allowing output to grow at least as rapidly as population. Similarly, continuing technical progress allows continuing output growth even if one factor is not increasing.

The price mechanism provides the correct incentives for these processes to occur. With a given supply of land, higher agricultural production raises the price of land and the rental paid for land. This provides an incentive to switch to less land-intensive production methods (heavy fertilizer usage, battery chickens) and an incentive to focus on technical progress that lets the economy get by with less land. A similar argument applies to any natural resource in finite supply.

Capital accumulation

Post-war theories of economic growth date back to work in the 1940s. In the late 1950s, Bob Solow of MIT assembled the nuts and bolts of neoclassical growth theory – the basis of empirical work ever since.¹ Along the steady-state path, output, capital and labour grow at the same rate. Hence output per worker and capital per worker are constant.

The theory is *neoclassical* because it does not ask how actual output gets to potential output. Over a long enough period, the only question of interest is what is happening to potential output itself. Neoclassical growth theory simply assumes that actual and potential output are equal.

Along the **steady-state path**, output, capital and labour grow at the same rate. Hence output per worker and capital per worker are constant.

In this long run, labour and capital grow. Usually, equilibrium means that things are not changing. Now we apply equilibrium not to levels but to growth rates and ratios. The **steady state** is the long-run equilibrium in growth theory.

In a growing economy, **capital widening** extends the existing capital per worker to new extra workers.

Capital deepening raises capital per worker for all workers.

Assume that labour grows at a constant rate n . To keep things simple, we also assume a constant fraction s of income is saved; the rest is consumed. Aggregate capital formation (public and private) is the part of

output not consumed (by both public and private sectors). Investment first **widens** and then perhaps **deepens** capital.

To keep capital per person constant, we need more investment per person the faster is population growth n (extra workers for more capital per person k that has to be provided. Figure 28.3 plots the line nk along which capital per person is constant. Any investment above this line implies capital deepening is taking place, whereas below this line capital per person must be falling.

Adding more capital per worker k increases output per worker y , but with diminishing returns: hence the curve y in Figure 28.3. Since a constant amount of output is saved, sy shows the saving per person. Since saving and investment are equal, it also shows investment per person.

In the steady state, capital per person is constant. Hence investment per person sy must equal nk , the investment per person needed to keep k constant by making capital grow as fast as labour. k^* is the steady-state capital per person and y^* the steady-state output per person. Capital, output and labour all grow at the same rate n along this steady-state path.

Figure 28.3 also shows what happens away from the steady state. If capital per worker is low, the economy is left of the steady state. Per capita saving and investment sy exceed nk , the per capita investment required to keep capital in line with growing labour. So capital per person rises. Conversely, to the right of the steady state, sy lies below nk and capital per person falls. Figure 28.3 shows that, from whatever level of capital the economy begins, it gradually converges to the (unique) steady state.

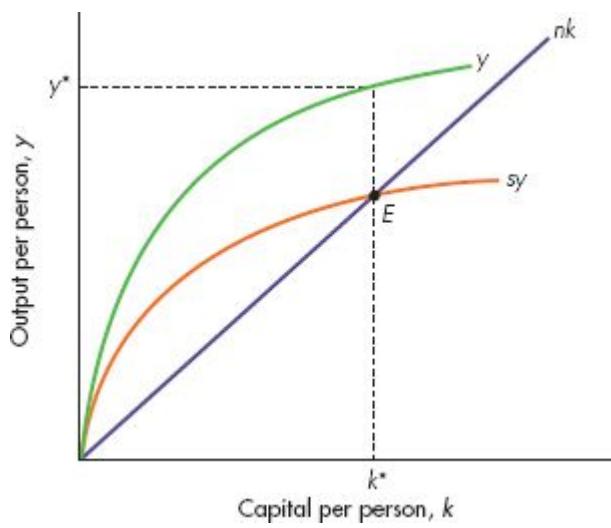
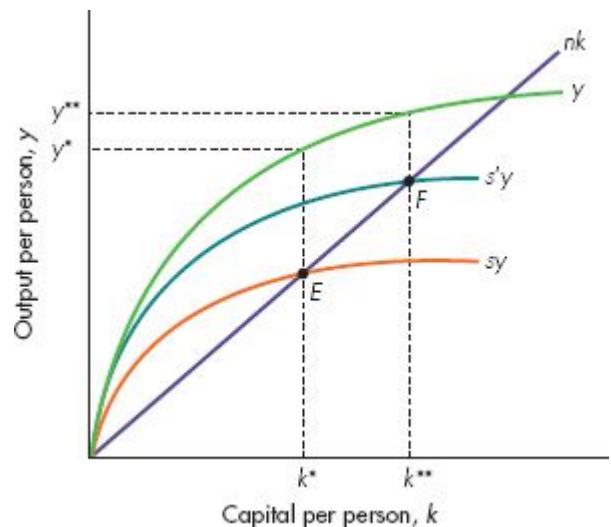


Figure 28.3 Neoclassical growth

The line nk shows the investment per person that maintains capital per person while labour grows. y shows output per person, and sy is both saving and investment per person. At the steady state E , investment is just sufficient to keep capital per person constant at k . Per capita output is then y^* . Output and capital grow with population.



The original steady state is at E . An increase in the fraction of income saved, from s to s' , leads to a steady state at F . This raises capital and output per worker, but eventually has no effect on the growth rate. Since y^{**} is constant, output and labour still grow at rate n .

Figure 28.4 A higher saving rate

We could augment Figure 28.3 by also recognizing that capital depreciates over time, becoming less and less productive. Suppose the rate of depreciation is d . In order to preserve the amount of capital per worker when the labour force is growing but inherited capital is depreciating, it now takes investment per worker of $(n - d)k$. Hence, to recognize depreciation, in Figure 28.3 we could replace the orange line nk by a steeper line $(n - d)k$. This will intersect sk at a lower level of capital per person k^* , implying also a lower level of output per person y^* . Faster depreciation increases the cost of keeping capital, and reduces steady state living standards.

A higher saving rate

Suppose people permanently increase the fraction of income saved, from s to s_9 . We get more saving, more investment and hence a faster rate of output growth. Oh no, we don't! Figure 28.4 explains why not.

There is no change in the production function relating output to inputs. At the original saving rate s , the steady state is at E . At the higher saving rate, s_9y shows saving and investment per person. At F it equals nk , the per capita investment needed to stop k rising or falling. Thus F is the new steady state.

F has more capital per worker than E . Productivity and output per worker are higher. That is the permanent effect of a higher saving rate. It affects levels, not growth rates. In *any* steady state, L , K and Y all grow at the same rate n , and that rate is determined ‘outside the model’: it is the rate of growth of labour and population. We return to this issue shortly.

In Figure 28.4 the higher saving rate raises output and capital per worker. To make the transition from E to F , there must be a temporary period in which capital grows faster than labour; only then can capital per worker rise as required. A higher saving rate, if successfully translated into higher investment to keep the economy at full employment, causes faster output growth for a while, but not for ever. Once capital per worker rises sufficiently, higher rates of saving and investment go entirely in capital widening, which is now more demanding than before. Further capital deepening, the basis of productivity growth, cannot continue without bound.

MATHS 28.1

NEOCLASSICAL GROWTH THEORY

Output per head is $f(k)$. With a constant saving rate s , saving per head is $sf(k)$.

In a simple economy with no government or foreign sector, at full capacity this must equal investment per head, which comprises replacement investment per head nk and capital deepening that adds to k . If \dot{k} denotes the rate of change of k ,

$$\dot{k} + nk = sf(k)$$

In the steady state k^* , the growth of k is zero, hence $nk^* = sf(k^*)$. Elsewhere,

$$\dot{k} = sf(k) - nk$$

Thus,

$$\dot{k} > 0 \quad \text{if } f(k)/k > n/s \quad \text{and} \quad \dot{k} < 0 \quad \text{if } f(k)/k < n/s$$

With diminishing returns to adding extra units of k , $f(k)$ increases less quickly than k itself. Hence, for $k < k^*$ it must be the case that $f(k)/k$ is less than n/s . Conversely, for $k > k^*$, $f(k)/k$ must exceed n/s . Thus, capital deepening is positive whenever k is less than its steady-state value k^* , and is negative whenever k exceeds its steady-state value k^* , confirming that the neoclassical economy converges to its unique steady state whatever level of k it begins with.

In this steady state, $nk^* = sf(k^*)$. Hence, for given n , an exogenous increase in the saving rate s must increase $k^*/f(k^*)$. Because the function $f(k)$ has diminishing returns to increasing k , it requires a higher value of k^* to increase $k^*/f(k^*)$ when s increases. Higher saving leads to a rise in the steady-state level of capital per head.

28.6 Growth through technical progress

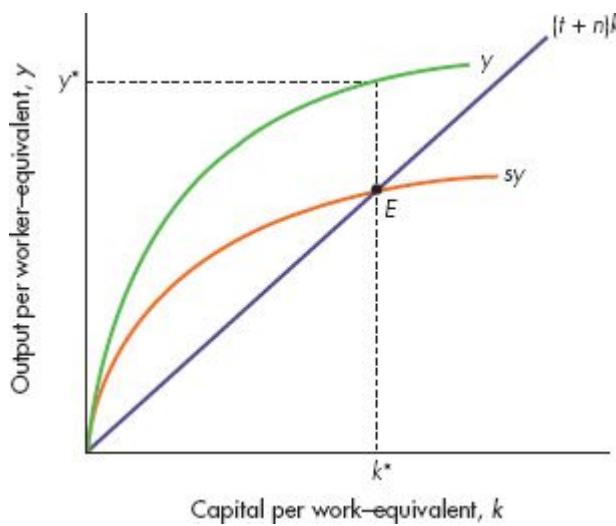
Labour-augmenting technical progress increases the effective labour supply.

We have made a lot of progress, but still have some problems. First, the theory does not fit *all* the facts. So far, the theory says output, labour and capital all grow at rate n . Although capital and output do grow at similar rates, in practice both grow more rapidly than labour. That is why we are better off than our great-grandparents.

The answer may lie in technical progress, which we ignored in trying to explain output growth entirely through growth in factor supplies (population growth and the accumulation of capital). **Labour-augmenting technical progress** provides the answer.

Population growth might eventually double the number of workers. Imagine instead that the number of workers is constant but that new knowledge allows the same workers to do the work of twice as many as before, as if the population had grown.

Suppose this progress occurs at rate t . Effective labour input grows at rate $(t + n)$ because of technical progress and population growth. Figure 28.5 resembles Figure 28.3 but replaces actual workers with worker-equivalents. To make this valid, we have to measure capital and output not per worker but per worker-equivalent. Worker-equivalents are created by population growth or technical progress. Otherwise the diagram is identical. Thus, investment at the rate $(t + n)k$ is now needed to ensure that capital per worker-equivalent remains constant.



The line $[t + n]k$ shows the investment per worker-equivalent to maintain capital per worker-equivalent when effective labour input grows, so shows saving (and thus investment) per worker-equivalent. At the steady state E, investment is just sufficient to keep capital per worker-equivalent constant at k . Output per worker-equivalent is constant at y . Output and worker-equivalents grow at the rate $[t + n]$, but actual people grow only at the rate n . Hence, output per actual person grows at the rate $t[t + n]$.

Figure 28.5 Neoclassical growth with technical progress

E remains the steady state. Output per worker-equivalent and capital per worker-equivalent are constant. Since worker-equivalents grow at rate $t + n$, so must capital and output. Since actual workers increase at rate n , output and capital per actual worker each increase at rate t . Now our growth theory fits all the facts. Living standards grow over time at rate t .

It is uncomfortable that the two key growth rates, n and t , are determined outside the model. For that reason, for the next 30 years the main use of this growth theory was in growth accounting: showing how to decompose actual output behaviour into the parts explained by changes in various inputs and the part residually explained by technical progress. We next examine the results of accounting for growth.

28.7 Growth in the OECD

The Organization for Economic Cooperation and Development is a club of the world's richest countries, from industrial giants like the US and Japan to smaller economies like New Zealand, Ireland and Turkey. Table 28.1 shows productivity growth of selected OECD countries since 1950.

Table 28.1 Average annual growth in real output per person employed (%)

	OECD	Japan	Germany	Italy	France	Sweden	UK	US
1950–73	3.6	8.0	5.6	5.8	4.5	3.4	3.6	2.2
1973–90	1.4	2.9	2.3	2.4	2.8	1.6	1.8	0.4
1990–2007	1.5	1.2	2.4	1.1	1.4	2.1	1.6	1.6
2008–11	0.5	0.0	20.2	20.7	0.0	0.5	20.6	1.4

Sources: S. Dowrick and D. Nguyen, 'OECD comparative economic growth 1950–85', *American Economic Review* 79 (1989): 1010–1030; OECD, *Economic Outlook*.

During the post-war boom years 1950–73, productivity grew strongly in a climate of rapid trade expansion, investment and recovery. These happy days ended in 1973 in all OECD countries. Several explanations were put forward. Some stressed the rise in trade union power, resulting in their enjoying greater legal protection in the 1970s. If this explanation had been correct, the supply-side reforms of the late 1980s and 1990s should have led to high productivity growth in the 1990s. They did not.

The first OPEC oil price shock, when real oil prices quadrupled, also occurred in 1973. This had two effects. First, it diverted R&D to long-term efforts to find alternative energy-saving technologies. These efforts take decades to pay off and raise actual productivity. Second, higher energy prices made much of the capital stock economically obsolete overnight. Energy-guzzling factories were closed. The world lost part of its capital stock, which reduced output per head. In practice, scrapping took a long time, and was given renewed impetus by another sharp rise in oil prices in 1980/81. That is why its effects were drawn out over such a long period, lasting for much of the 1980s.

ACTIVITY 28.1

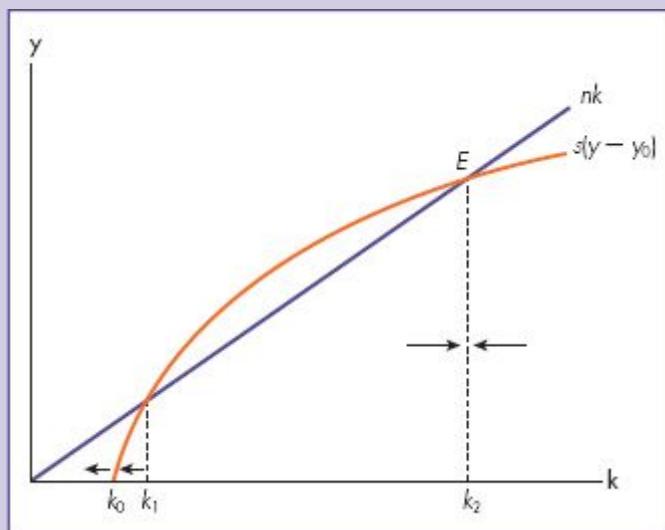
ABORTED TAKE-OFFS ON THE GROWTH RUNWAY

We assume people save a constant fraction s of their income. Even poor people earning only y save sy and consume $(1 - s)y$. But if y is low enough, $(1 - s)y$ is too low to stop starvation. So they consume all their income and save none. Below a critical income level y_0 , saving is zero. What does the Solow diagram look like now? Look at the diagram below. Suppose k_0 is the capital per person that just generates the critical income y_0 . Higher capital generates saving as in previous diagrams, and nk is still the gross investment needed to maintain a given capital-labour ratio in the face of a growing population. There are now three steady states!

If capital begins above k_1 , the economy converges to the steady state at E . Between k_1 and k_2 , saving and investment exceed the amount needed for capital-widening; capital-deepening also occurs and the economy grows. Above k_2 , saving and investment are insufficient to maintain the capital-labour ratio, and the economy

shrinks. Either way, it ends at E . This is the case analysed in Figure 28.2 and Figure 28.3. Suppose, next, the economy begins at exactly k_1 . Saving and investment just maintain the capital–labour ratio. So this is a steady state, but an unstable one. A little above k_1 the economy begins converging on E . And below k_1 there is insufficient saving and investment to provide for the growing population. Capital per person shrinks and keeps shrinking until the economy reaches k_0 .

In this model, countries beginning with capital below k_1 are stuck in a poverty trap. They cannot break out. All output is consumed to prevent starvation. There is never a surplus to begin accumulation and growth. This model can also explain why convergence seems to occur within the OECD (countries already above k_1), but why simultaneously many countries are stuck in poverty. Modern growth in the last two centuries began when some key events first generated the surplus to allow saving and accumulation to begin.



Questions

- (a) Why is there no poverty trap when saving is proportional to income?
- (b) When a poverty trap exists, is the payoff to overseas aid from rich countries greater if it is concentrated on helping poor countries break out of the poverty trap?
- (c) The poverty trap shown above is based on there being a minimum level of per capita consumption. Could we get a

poverty trap based on different population growth rates above and below some critical threshold of living standards? Is this plausible?

To check your answers to these questions, go to page 682.

Neither the Internet boom nor supply-side reforms restored the productivity growth rates that the rich countries enjoyed prior to 1973. Emerging market economies, such as China, India, Brazil and Russia, are now where the action is. We turn to their story in Part Five. For rich mature economies, Table 28.1 confirms that underlying productivity growth showed a very modest improvement after 1990, until it was dramatically interrupted by recession after the financial crash.

Why was productivity growth low or negative during 2008–11? OECD countries did not mysteriously forget the skills they had learned over previous decades, nor did their capital explode. Table 28.1 is alerting us to a short-term effect, the normal fluctuation of productivity with output over the business cycles, discussed in the previous chapter. If normal output growth is resumed, we expect to see productivity growth return to normal levels again.

Having discussed differences in growth across periods, we now examine differences across countries. The one sheds light on the other. The fact that OECD countries move together across sub-periods shows that many aspects of growth are outside a country's own control. Technical progress diffuses across countries quite quickly, wherever it originates. Countries are increasingly dependent on the same global economy.

Even so, growth rates differ markedly across countries. Can growth theory explain why? First, it suggests that, if countries have access to the same technology, differences in output growth should reflect differences in labour force growth. Table 28.1 provides some degree of corroborating evidence: differences in per capita output growth are less marked than differences in output growth.

Second, we need to know how long it takes to get to the steady state, a question to which Figures 28.3 and Figures 28.5 provide no direct answer. Is output growth over two or three decades an adjustment towards the steady state or can we assume that an economy has reached it within that time?

The convergence hypothesis

The **convergence hypothesis** asserts that poor countries grow more quickly than average and rich countries grow more slowly than average.

Figures 28.3 and Figure 28.5 have a unique steady state at E . Whatever the level of capital per worker with which an economy begins, the economy eventually converges to E . Poor countries with low initial capital grow extra rapidly until they reach the steady-state growth rate of output and capital; rich countries with a very high inheritance of capital grow at below-average rates until capital per worker falls back to its steady-state level k^* .

When capital per worker is low, it does not take much investment to equip new workers with capital (capital-widening), so the rest of investment can go on raising capital per worker (capital-deepening). When capital per worker is already high, it takes a lot of saving and investment just to maintain capital-widening, let alone to deepen capital. This is one reason for the convergence hypothesis.

This explanation for convergence relies purely on the effect of capital accumulation. A second explanation for convergence or ‘catch-up’ operates through a different channel. Technical progress no longer falls out of the sky at a fixed rate. Suppose instead we have to invest real resources (universities, research labs, R&D) in trying to make technical improvements. It is rich countries that have the human and physical capital to undertake these activities, and it is in rich countries that technical progress is made. However, once discovered, new ideas are soon disseminated to other countries.

Since poorer countries do not have to use their own resources to make technical breakthroughs, they can devote their scarce investment resources to other uses, such as building machines. By slipstreaming richer countries, they can temporarily grow faster.

CASE 28.2

Does convergence occur in practice?

The table below shows World Bank estimates of per capita income in 1987, 2008 and the ratio of 2008 to 1987. East Asian economies such as China and South Korea grew very quickly during the last 30 years. India (not shown below) is also now growing strongly. Yet convergence cannot be a powerful force in the world or the very poorest countries would all be growing very rapidly. In reality, many poor countries stay poor and sometimes even decline in absolute terms.

Within the rich OECD countries, convergence is much more reliable. The richest OECD countries tend to grow less quickly than the poorer OECD countries.

Why did the East Asian ‘tigers’ grow so quickly in the post-war period? What was their secret? Professor Alwyn Young² of MIT has shown that there is little mystery about their rapid growth, even though they did sustain dramatic rates. These economies managed rapid growth in measured inputs – labour (via increases in participation rates), capital (via high saving and investment rates) and human capital (via substantial expenditure on education). Once we allow for the rapid growth of these inputs, Young showed that the growth of output in the tigers was not very different from what standard estimates, based on OECD and Latin American countries, would have led us to expect.

Generally, growth seems to be fostered by two conditions: absence of internal strife and openness to the world economy. Once China put insularity and the Cultural Revolution behind it, the potential for catching up was enormous. India had less internal strife, but took off only after it embraced the world economy and relaxed its more bureaucratic controls. Civil war held back Nigeria despite its oil wealth. Indeed, there is considerable evidence that mineral-rich countries without a long tradition of stable government suffer disproportionate incidence of civil war – fighting for the spoils – to the detriment of economic growth and higher living standards.

Note finally that Switzerland, with much the highest living standard, has one of the slowest rates of growth of per capita GNP. The Swiss are rich today because they were rich yesterday, a secret that they discovered long ago.

Per capita GNP (2012 US\$000s)

	Annual real growth (%)		
	GDP	Per capita GDP	Population
Vietnam	6.8	6.0	0.8
India	5.8	5.0	0.8
Bangladesh	5.1	3.9	1.2
China	4.7	4.6	0.1
Malaysia	4.3	3.3	1.0
Turkey	4.1	3.4	0.7
Mexico	3.7	3.2	0.5
Russia	2.5	3.2	-0.6
Poland	2.1	2.7	-0.5

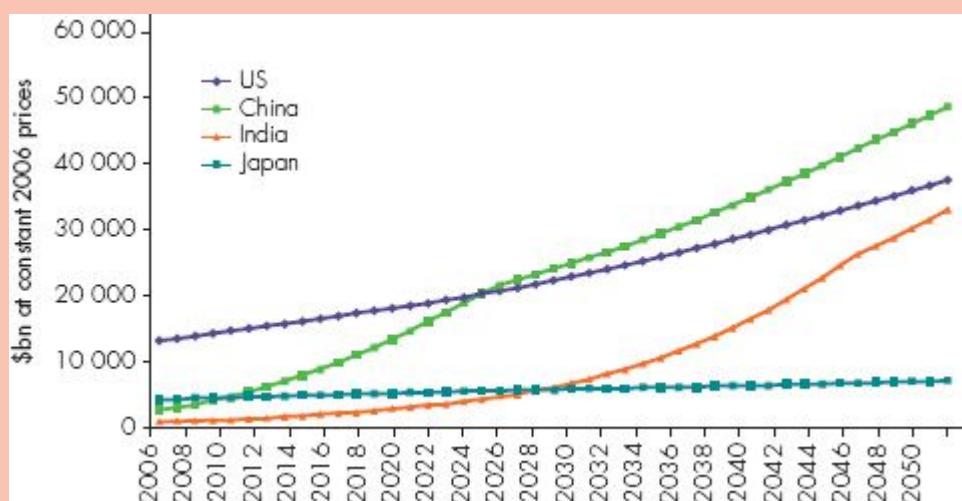
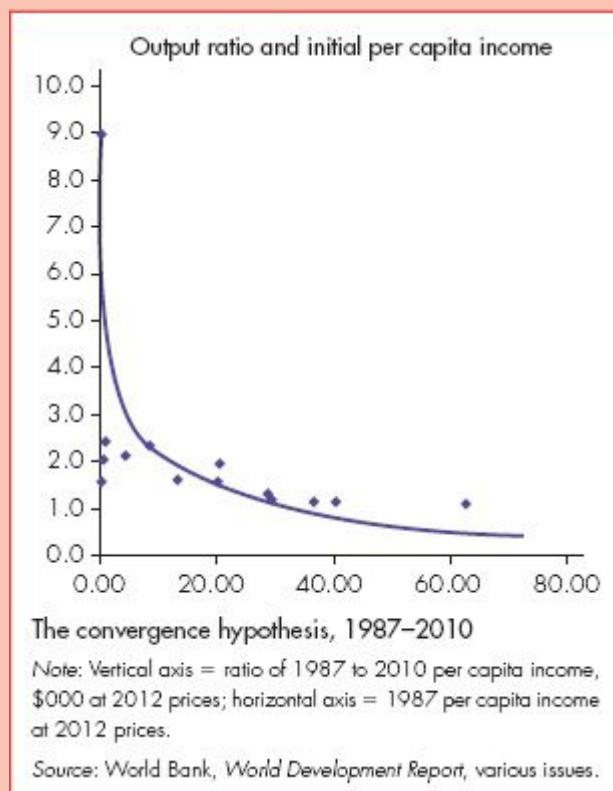
Source: World Bank, *World Development Report*, various issues.

The figure opposite plots the final column, the ratio of per capita income in 2010 relative to 1987, on the horizontal axis, and on the vertical axis plots 1987 per capita income in ten-thousands of US dollars, obtained by dividing the third column of the table by ten in order to keep the two scales comparable. The figure conveys two messages. First, on average richer countries grow more slowly. Second, individual country performance can depart significantly from this underlying relationship.

With this understanding, what should we expect for the next 40 years, long after the consequences of the financial crash have worked themselves out? Global consultancy PricewaterhouseCoopers (PwC) makes brave projections for the future, based largely on the framework we have set out. It estimates population growth, the evolution of skills and human capital, investment in physical capital, and rates of technical progress and its dissemination across countries. From this information, it makes estimates of future growth in GDP.

The *World in 2050* chart, drawn from PwC's 2008 report, refers to the projected evolution of aggregate GDP. Who will be the economic superpowers in 2050? At present, US GDP significantly exceeds that of the second-largest economy, Japan. By 2025, China will have overhauled the US, and by 2050 will have a significant economic lead. With its much larger population, catch-up in productivity is all that is required. With an aged population, Japan will fail to keep up: old people consume but do not produce, and so

attract resources away from investment and accumulation. After a slower beginning, India becomes the most exciting story of all. With the fastest population growth and the second-largest population to start with, India should start to narrow the gap on China and will overtake the US eventually, but not before 2050.



The table below shows PwC estimates for a variety of emerging markets during 2007–50, isolating the effect of population growth as well as the general catch-up in productivity levels implied by more rapid productivity growth in poorer countries.

	Annual real growth (%)		
GDP	Per capita GDP	Population	
Vietnam	6.8	6.0	0.8
India	5.8	5.0	0.8
Bangladesh	5.1	3.9	1.2
China	4.7	4.6	0.1
Malaysia	4.3	3.3	1.0
Turkey	4.1	3.4	0.7
Mexico	3.7	3.2	0.5
Russia	2.5	3.2	20.6
Poland	2.1	2.7	20.5
<i>Source:</i> PwC, <i>The World in 2050</i> .			

The poorer countries have more rapid GDP growth not merely because they have opportunities for productivity catch-up but also because they often have more rapid population growth, except in China with its one-child per family policy. Whether this will continue until 2050, as assumed in the table, is hard to assess at this juncture. The middle-income countries (Malaysia through to Poland) are expected to have fewer opportunities for rapid productivity growth and, in the case of Eastern Europe and Russia, may actually experience falling populations. This helps their per capita growth – capital-widening is less of a burden – but not their aggregate GDP growth.

In a path-breaking empirical study over 50 years ago, Professor Bob Solow compared growth rates across countries and across time, documented how much growth could be traced to growth of inputs of labour and capital via a standard production function, and attributed the unexplained part of economic growth to unmeasured technical progress.

Case 28.2 already provides two alarm bells. First, countries do not immediately share the same technical knowledge, providing scope for catch-up by poorer countries. Second, even allowing for this, different countries behave differently. The part of output growth not explained by the growth of measured inputs is known as the Solow residual.

The **Solow residual** measured our initial ignorance about economic growth, and economists have spent the last 50 years trying to explain more and more of it. Two useful approaches have been to keep explicit track of more inputs – such as energy and knowledge capital – and to elaborate a dynamic model of technical progress in which some countries get new information before others.

The part of output growth not explained by the growth of measured inputs is known as the **Solow residual**.

One early attempt, by Professor Nick Crafts of Warwick University, took the Solow residuals and tried to see how much of them could then be explained by catch-up. The lower a country's per capita GDP relative to that of the US (the assumed technical leader), the larger should be the potential for catch-up.

Crafts discovered that there is a systematic role for catch-up, as we would have expected from Case 28.2, but also that, after allowing for 'average catch-up' for a country of that living standard, big differences remain across countries. These may reflect the social and political framework in which the economy must operate. Change usually helps the majority but has very adverse effects on a few people whose skills are made obsolete or whose power is suddenly removed. The large number of winners should club together to buy off the few big losers, allowing change to proceed. Some societies are much better than others at organizing the deals that allow catch-up to be achieved more rapidly.

Increasingly, the effort of research economists is now focused on analysing and measuring the accumulation of knowledge capital. Amazingly, recent estimates suggest that investment in knowledge creation is even more important than investment in physical capital in generating high levels of GDP. With around 80 per cent of GDP now comprising the supply of services rather than the production of goods, perhaps this should come as no surprise. Know-how matters a lot.

This also helps explain some historical puzzles, such as why Germany and Japan grew so quickly after the Second World War. Part of the modern answer is that their advanced knowledge survived intact even if the bombs had destroyed their buildings. Conversely, if foreign aid is to help some of the world's poorest countries, supplying food and shelter may not be enough. Education and training are hugely important.

28.8 Endogenous growth

Endogenous growth implies that the steady-state growth rate is affected by economic behaviour and economic policy.

In its simplest form, Solow's theory makes economic growth depend on population growth and technical progress. Both proceed at given rates. The subsequent literature on catch-up makes technical progress respond to economic circumstances and the political and cultural environment. It would be nice to have a stronger link between economic behaviour and the rate of economic growth. We want to make growth *endogenous*; that is, determined within our theory.

The original insight is due to Professor Paul Romer. Saving, investment and capital accumulation lie at the heart of growth. In Solow's theory, applying more and more capital to a given path for population runs into the diminishing marginal product of capital. It cannot be the source of permanent growth in productivity.

We know there must be diminishing returns to capital alone at the level of individual firms; otherwise one firm would get more and more capital, become steadily more productive and gradually take over the entire world! Because diminishing returns to capital hold at the level of the firm, economists had assumed they held also at the level of the economy.

Romer's insight was the possibility (likelihood?) that there are significant externalities to capital. Higher capital in one firm increases productivity in *other* firms. When BT invests in better equipment, other firms can do things previously impossible. The Internet and the iPad have similar benefits for other technologies. The insight also applies to human capital. Training by one firm has beneficial externalities for others.

Thus the production function of each individual firm exhibits diminishing returns to its own capital input, but also depends on the capital of other firms. No firm, acting in isolation, would wish to raise its capital without limit. But when all firms expand together, the economy as a whole may face constant returns to aggregate capital.

Consider the following simple example of the aggregate economy. Per capita output y is proportional to capital per person k . To isolate the role

of accumulation, suppose there is no technical progress. Thus $y \leq Ak$, where A is constant, and there are constant returns to accumulating more capital. Given a constant saving rate s and population growth at rate n , is there a steady state in which capital per person grows at rate g ? If so, investment for capital deepening is gk and investment for capital-widening, to keep up with population growth, is nk . Hence, in per capita terms:

$$\text{Gross investment} = (g + n)k = sy = sAk = \text{gross saving}$$

Hence $gk = (sA - n)k$ and the steady-state growth rate g is

$$g = (sA - n)$$

Why does this confirm the possibility of *endogenous* growth? Growth now depends on parameters that can be influenced by private behaviour or public policy. In the Solow model, without technical progress, steady-state growth is always n , whatever the saving rate s or the level of productivity A . Equation (2) says that any policy that succeeded in raising the saving rate s would *permanently* increase the *growth rate* g . Similarly, any policy achieving a one-off rise in the *level* of A , for example greater workplace efficiency, would permanently increase the growth rate of k . Since $y \leq Ak$, this means permanently faster output growth.

Not only can government policy affect growth in this framework, government intervention may increase efficiency. In the simple Romer model outlined above, there are externalities in capital accumulation: individual firms neglect the fact that, in raising their own capital, they also increase the productivity of *other* firms—capital. Government subsidies to investment might offset this externality.

Since Romer's original work there has been huge interest in endogenous growth. Sustaining small additions to annual growth rates eventually makes a big difference to living standards. As a result of this research we now have many potential channels of endogenous growth. For example, instead of assuming that the rate of technical progress is given, we can model the industry that undertakes R&D to produce technical progress. Constant returns in this industry will generate endogenous growth. In fact, constant returns to aggregate production of any *accumulable* factor (knowledge, capital, and so on) will suffice.

Note, too, that endogenous growth models explain why growth rates in different countries might permanently be different. This might explain why convergence does not take place and why some countries remain poor indefinitely. Different countries have different growth rates g .

While endogenous growth theory is an exciting development, it also has its critics. Most criticisms boil down to a key point. Whatever the relevant accumulatable factor, why should there be *exactly* constant returns in the aggregate? With diminishing returns, we are back in the Solow model where long-run growth is exogenous. With increasing returns, the economy would settle not on steady growth but on ever more rapid expansion of output and capital. We know this is not occurring. So for endogenous growth theory to be the answer, only constant returns to accumulation will do. Some people think this seems just too good to be true.

28.9 The costs of growth

Can the benefits of economic growth be outweighed by its costs? Pollution, climate change, congestion and a hectic lifestyle are a high price to pay for more cars, washing machines and video games.

Since GNP is an imperfect measure of the true economic value of the goods and services produced by the economy, there is no presumption we should want to maximize the growth of measured GNP. We discussed issues such as pollution in Part Three. Without government intervention, a free market economy produces too much pollution. But complete elimination of pollution is also wasteful. Society should undertake activities accompanied by pollution up to the point at which the net marginal benefit of the goods produced equals the marginal pollution cost imposed on society. Government intervention, through pollution taxes or regulation of environmental standards, can move the economy towards an efficient allocation of resources in which marginal social costs and benefits are equalized.

The **zero-growth** proposal argues that, because higher measured GNP imposes environmental costs, it is best to aim for zero growth of measured GNP.

The full implementation of such a policy would (optimally) reduce the growth of measured GNP below the rate where there is no restriction on activities such as pollution and congestion. And this is the most sensible way in which to approach the problem. It tackles the issue directly. In contrast, the **zero-growth** solution is a blunt instrument.

The zero-growth approach fails to distinguish between measured outputs accompanied by social costs and measured outputs without additional social costs. It does not provide the correct incentives. The principle of targeting, a key insight of the welfare economics discussed in Part Three, suggests that it is more efficient to tackle a problem directly than to adopt an indirect approach that distorts other aspects of production or consumption. Thus, when there is too much pollution, congestion, environmental damage or stress, the best solution is to provide incentives that directly reduce these phenomena. Restricting growth in measured output is a crude alternative, distinctly second-best.

Some problems might evaporate if economists and statisticians could measure true GNP more accurately, including the ‘quality of life’ activities (clean air, environmental beauty, sustainable climate, and so on) that yield consumption benefits but at present are omitted from measured GNP. Voters and commentators assess government performance against measurable statistics. A better measure of GNP might remove perceived conflicts between measured output and the quality of life.

This is also a good way to address ‘sustainable growth’. At present, Mediterranean beauty spots become concrete jungles of hotels and bars; once the environment is spoiled, upmarket tourists move on to somewhere else. An economist’s advice, however, is not to abandon being a tourist destination, but to keep track of environmental depreciation and only engage in activities that show a clear return after proper costing of environmental and other damage. Embodying these costings in actual charges also provides the market incentive to look after the environment.

This also provides the answer to those who argue that tackling climate change will hamper economic growth. Just as we want congestion charging to *reduce* some outputs (rush-hour traffic), we want environmental pricing to *reduce* some activities (greenhouse gas emissions, lax building insulation). In both cases, the objective is to get aggregate output, *properly measured*, to increase!

No matter how complete the framework, the assessment of the desirable growth rate will always be a normative question hinging on the value judgements of the assessor. Switching resources from consumption, however defined, to investment will nearly always reduce the welfare of people today but allow greater welfare for people tomorrow. Nowhere is this clearer than in the speed with which we try to deal with climate change. More sacrifice today will make life easier tomorrow; less sacrifice today will compound the problems for our children's children. The priority attached to satisfying wants of people at different points in time is always a value judgement.

Summary

- Supply-side economics analyses one-off changes in aggregate supply and potential output. These arise from changing the quantity of labour employed or changing the productivity of those employed.
- For a given population, the quantity of labour employed is affected by labour force participation, by incentives of the labour force to accept a job and by incentives about how long and how hard to work. Changes in incentives have income and substitution effects, often operating in different directions, somewhat mitigating their overall effectiveness.
- Productivity can be enhanced through greater competition, lighter regulation (within limits) and allowing obsolete businesses to cease (especially if accompanied by effective retraining and relocation of the workforce).
- The possibility of **hysteresis** means that temporary recessions may have permanently adverse effects on potential output. If so, demand stabilization may contribute to good long-run supply performance.
- Economic growth is the percentage annual increase in real GDP or per capita real GDP in the long run. It is an imperfect measure of the rate of increase of economic well-being.
- Measured GDP omits the value of leisure and of untraded goods and bads that have an impact on the quality of life. Differences in income

distribution make per capita real GDP a shaky basis for comparisons of the welfare of the typical individual in different countries.

- Significant rates of **growth of per capita GDP** occurred only in the last two centuries in the advanced economies. In other countries persistent growth is even more recent.
- Potential output can be increased either by increasing the inputs of land, labour, capital and raw materials, or by increasing the output obtained from given input quantities. **Technical advances** are an important source of productivity gains.
- An apparently **fixed supply of a production input**, such as a particular raw material, need not make growth impossible in the long run. As the input becomes scarce, its price rises. This makes producers substitute other inputs, increases incentives to discover new supplies and encourages inventions that economize on the use of that resource.
- The simplest theory of growth has a **steady state** in which capital, output and labour all grow at the same rate. Whatever its initial level of capital, the economy converges on this steady-state path. This theory can explain output growth but not productivity growth.
- Labour-augmenting technical progress allows permanent growth of labour productivity and enables the simple growth theory to fit many of the facts.
- There is a **tendency of economies to converge**, both because **capital deepening** is easier when capital per worker is low and because of **catch-up in technology**. Implementing technical change may depend on how well society is organized to buy off (or defeat) the losers.
- **endogenous growth** are built on constant returns to accumulation. If aggregate investment does not encounter diminishing returns to capital, choices about saving and investment can affect the long-run growth rate of productivity. An externality on a giant scale provides a powerful rationale for government intervention to encourage education, training and physical capital formation.

- Nevertheless, endogenous growth rests on the presence of constant returns to accumulation. Nobody has yet explained why this should hold.

Review questions



EASY

- 1 (a) What is the distinction between total output and per capita output? Which grows more rapidly? Why? Does it always grow? (b) What are the two channels by which potential output can be increased using supply-side policies?
- 2 Name two economic bads. Suggest feasible ways in which they might be measured. Should they be included in GNP? Could they be?
- 3 Choose the correct answer: Countries that isolate themselves from the world economy tend to grow slowly because (a) they fail to learn about technical progress elsewhere, (b) without competition, they have insufficient incentive to invest, (c) there are other adverse consequences of the political regime that took such a decision, (d) all of the above, (e) none of the above.
- 4 **Common fallacies** Why are these statements wrong? (a) Since the earth's resources are limited, growth cannot continue forever. (b) If we save more, we'd definitely grow faster.
- 5 Choose the correct answer: The empirical correlation between countries that possess extreme mineral wealth and the prevalence of civil wars suggests that: (a) wars raise the demand for resources and encourage exploration for minerals, (b) when easy wealth is available, it increases the incentive to fight over the spoils provided by nature rather than co-operate to produce goods and services, (c) mineral wealth attracts foreign predators.
- 6 'Because we know Malthus got it wrong, we are relaxed about the fact that some minerals are in finite supply.— Is there a connection? Explain.
- 7 Several decades ago, China adopted a policy of a maximum of one child per family. Using the analysis of this chapter, explain what the purpose of this might have been. Illustrate in a diagram.
- 8 Suppose the private sector has a given saving rate out of disposable income, but that the public sector levies taxes and utilizes all of this revenue for investment. If the private and public sector are equally

efficient, what happens to the long-run growth rate? What happens to per capita incomes eventually?

MEDIUM

- | 9 Consider an economy in which there is constant population. Each firm's production function exhibits diminishing returns to its own capital accumulation. However, each firm creates beneficial production externalities for other firms. In the aggregate, the economy faces constant returns to capital accumulation, so that $y \leq f(k)$. If the saving rate is 0.2, what is the permanent rate of growth of capital and output?
- | 10 Can technical progress be negative?
- | 11 'If the convergence hypothesis is correct, the poor African countries should have grown long ago!— Is this correct? Do newer approaches to economic growth help explain why some countries remain so poor?
- | 12 'Britain produces too many scientists, too few engineers.— What kind of evidence might help you decide if this is true? Will a free market lead people to choose the career that most benefits society?

HARD

- | 13 Consider a planet in which population grows at the constant rate n and people save a constant fraction s of their per capita output. Output is produced by environmental capital k , which depreciates at a constant d . Gross investment is used only to improve environmental capital, and $y \leq f(k)$ so that output depends on environmental capital and there are diminishing returns to environmental capital. (a) Illustrate the above information in a diagram. (b) What happens to gross investment per person? (c) Suppose the rate of environmental depreciation rises. What happens to the steady state level of output per person y^* ? (d) Is it true that if recycling were to reduce environmental depreciation, it would raise output per person in the long run?
- | 14 **Essay question** Is it in a country's best interests to focus on economic growth? Or, in other words, is growth good?

- 1 Solow won a Nobel Prize for his work on long-run growth. He is also famous for his one-liners. Since, in short-run analysis, he is an unrepentant Keynesian, many of his famous barbs are aimed at those who believe that prices clear markets quickly: 'Will the olive, unassisted, always settle half way up the martini?'
- 2 'A. Young, –The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience, Quarterly Journal of Economics 110 (1995): 641–680.

PART FIVE

The world economy

Part Five focuses on the world as a whole. What determines the pattern of international trade and the tariff policies pursued by individual countries? Can free trade benefit everyone? How can Europe compete with China and India? What trade policies should rich countries adopt?

Contents

29 Economics and the economy

CHAPTER 29

International trade

Learning Outcomes

By the end of this chapter, you should be able to:

- 1 analyse patterns of international trade
- 2 understand comparative advantage and the gains from trade
- 3 describe determinants of comparative advantage
- 4 understand why two-way trade occurs for the same product
- 5 describe trade policy
- 6 understand the principle of targeting
- 7 recognise motives for tariffs

International trade is part of daily life. Britons drink French wine, Americans drive Japanese cars and Russians eat American wheat. China makes European clothes but buys up raw materials that Europeans would otherwise have bought. There are three reasons why trade between the UK and Japan is different from trade between London and Birmingham.

First, international trade crosses national frontiers: governments can monitor this trade and treat it differently. It is hard to tax or regulate goods moving within a country and much easier to impose taxes or quota restrictions on goods imported from Taiwan or Japan. Governments have to decide whether or not such policies are desirable.

Second, if internal trade redistributes income between residents of that country, its government can always use the national tax system to offset this effect. If trade between Ireland and China leaves Ireland worse off, Irish fiscal policy cannot compensate in the same way.

Third, international trade may involve the use of different national currencies. An Italian buyer of American wine pays in euros but the American vineyard worker is paid in dollars. International trade involves international payments. We examined the system of international payments in Chapter 26.

This chapter concentrates on trade flows and trade policy. Who trades with whom, and in what commodities? Why does international trade take place? Countries trade with one another because they can buy foreign goods at a lower price than it costs to make the same goods at home.

Is this possible for all countries? International trade reflects *exchange* and *specialization*. International differences in the availability of raw materials and other factors of production lead to international differences in production costs and goods prices. Through international exchange, countries supply the world economy with the commodities that they produce relatively cheaply and demand from the world economy the goods made relatively cheaply elsewhere.

These benefits from trade are reinforced if there are scale economies in production. Instead of each country having a lot of small producers, different countries concentrate on different things and everyone can benefit from the cost reductions that ensue.

We discuss the benefits from international trade and examine whether our analysis can explain the trade flows that actually take place. There are many circumstances in which international trade can make countries better off, but trade can also carry costs, especially in the short run. Cheap foreign cars are great for UK consumers but less good for unemployed UK car workers.

Because foreign competition makes life difficult for some voters, governments are frequently under pressure to restrict imports. We conclude the chapter by discussing trade or commercial policy and whether it is a good idea to restrict imports.

29.1 Trade patterns

Every international transaction has both a buyer and a seller. One country's imports are another country's exports. To measure how much trade occurs, we can measure the total value of exports by all countries

or the total value of imports. Table 29.1 shows the value of world exports and, as a benchmark, the value relative to GNP in the world's largest single economy, the US.

Table 29.1 The value of world exports

World exports	1928	1935	1950	1973	2011
2012 £bn	446	187	281	1496	18000
As % of US GNP	57	27	20	40	115

Sources: League of Nations, *Europe's Trade*, Geneva, 1941; IMF, *International Financial Statistics*; *National Income Accounts of the United States*, 1928–49.

Table 29.2 Exports as a percentage of GDP

	1967	2011
Belgium	36	74
Netherlands	43	73
UK	18	27
France	14	30
Italy	17	25
US	5	14
Japan	10	10

World trade was destroyed in the 1930s by a tariff war and a series of competitive devaluations that eventually benefited nobody. After 1950, world trade grew rapidly, at an average annual rate of 8 per cent. International trade becomes ever more important to national economies. Between 1950 and 2011, exports as a fraction of GNP rose substantially in most countries. Details for selected countries are shown in Table 29.2.

As trade grew, both in absolute terms and relative to the size of national economies, the interdependence of national economies increased. Except for the largest countries (the US, Japan) countries are now very open to international trade. Trade between Paris and Brussels is *international* trade, but trade between New York and California is not. Events in other countries affect our daily lives much more than they did 20 years ago. We now look at who trades with whom.

World trade patterns

Table 29.3 shows the pattern of trade, which used to be dominated by the rich countries of Europe, North America and Japan. As late as 1980, the developed countries were the origin of and destination for 71 percent of world exports, most of this trade being among themselves. Trade between other countries – the poor and middle-income countries, and the former communist economies – accounted for only 8 per cent of world trade.

Table 29.3 Trade patterns, 2010 (% of world exports)

Origin of exports	Destination of exports	
	Europe and North America	Other
Europe and North America	31	25
Other	11	33

Source: www.wto.org.

The rapid growth of emerging market economies – particularly Less-developed countries (LDCs) have low per capita incomes. China, India and Brazil, but also the Asian ‘tigers’ (including Singapore, Malaysia, Thailand, South Korea and the Philippines) – and the economic liberalization of the former Soviet Union has led to rapid growth in their international trade. By 2010, the picture had changed substantially. Trade among developed countries, although absolutely much larger, had fallen as a share of world trade, whereas trade among the other countries had risen substantially. The **less-developed countries (LDCs)** as a whole range from the very poor, such as Bangladesh, to the nearly rich, such as Brazil. Having enjoyed per capita income growth of nearly 10 per cent a year for nearly 30 years, China is breaking all records in making the transition from very poor to nearly rich.

Less-developed countries (LDCs) have low per capita incomes.

The commodity composition of trade

In rich countries, services account for most of value added or GDP. International trade in services is growing rapidly, but from a small baseline. Trade in goods – or merchandise trade – remains important

because many countries import goods, add a little value and then re-export them. The value added makes a small contribution to GDP but gross flows of imports and exports of goods are large. By importing goods, adding a little value and re-exporting, it is even possible that the value of exports exceeds the value of GDP itself.

Table 29.4 distinguishes between *primary commodities* (agricultural commodities, minerals and fuels), manufactured or processed commodities (chemicals, steel and cars) and commercial services (banking, travel, insurance, tourism). Services were a negligible component of cross-border international trade in 1955, but have increased to nearly a quarter of all trade. Merchandise trade – trade in goods rather than services – has increased dramatically in absolute terms, but fallen as a share of total world trade.

Table 29.4 The composition of world exports

% share of	1955	2010
Commercial services	3	23
Merchandise trade (goods)	97	77
<i>of which</i>		
Primary commodities	49	42
Manufactures	48	35

Sources: GATT, *Networks of World Trade, 1955–76*; www.wto.org

Within merchandise trade, trade in primary products (food and agriculture, fuels and minerals) has remained fairly stable as a share of overall trade.

World trade: the issues

Tables 29.1 to 29.4 set out the facts. World trade has grown faster than world income, and is increasingly important. Thirty per cent of all international trade is trade between rich countries in Europe and North America, and this rises to nearly 40 per cent once we also include Japan. These rich industrial countries are the main export markets for emerging market economies, although emerging Asian economies trade increasingly among themselves. These facts help explain some of the key issues in world trade.

Raw material prices

For decades, less-developed countries – which tended to specialize in export of primary products – claimed that industrial countries exploited them by buying raw materials at a low price, in exchange for manufactured goods at a much higher price. Producers of coffee, sugar, copper and many other products would have liked to be able to copy OPEC and triple the price of their primary products without suffering a significant reduction in the quantities demanded.

The rise of China and India changed this picture. Their thirst for primary products – particularly raw materials but also food – substantially bid up the world price of primary products. China is now investing in the development of Africa to secure its raw material supplies in the long run. Australia and Brazil have become rich by exporting large quantities of raw materials at high prices. Conversely, countries dependent on imports of raw materials have faced an adverse supply shock.

Agricultural protection

Farmers in rich countries not only receive agricultural subsidies, for example via the EU Common Agricultural Policy, but also enjoy protection behind high tariffs on imported agricultural goods. Emerging market producers complain that exclusion of their exports from the richest markets not only reduces the quantity of what they can sell but also forces down the price when all their supply must be absorbed in the remaining world markets to which they have access. Reducing protection of farmers in rich countries is probably the single greatest contribution of assistance that the rich could make to the poor.

Manufactured exports from emerging market economies

These countries want to make their own manufactured goods and export them to the industrial countries. Brazil, Mexico and South Korea have major manufacturing industries, China is now a powerhouse, and Indian companies such as Tata are moving into Western markets. Exports to industrial countries led to complaints in industrial countries that jobs are threatened by competition from cheap foreign labour. Should Asian exports be restricted to prevent massive job losses in Western Europe and

North America or should rich countries take advantage of low costs in Asia?

Globalization

Lower transport costs and better information technology are gradually breaking down the segmentation of national markets and increasing competition between countries. This trend has been reinforced by reductions in tariffs as a matter of policy. Sometimes the pace of change has been rapid.

However, poor countries feel that the process is largely dictated by rich countries according to their own self-interest. Poorer countries feel pressurized to dismantle their own tariffs and allow in foreign investors, while rich countries remain reluctant to pay attention to concerns of poor countries.

Before examining these issues, we explain why international trade takes place at all.

29.2 Comparative advantage

The **opportunity cost** of a good is the quantity of other goods sacrificed to make another unit of that good.

Trade is beneficial when there are international differences in the **opportunity cost** of goods.

Suppose a closed economy with given resources can make DVD recorders or shirts. The more resources used to make DVD recorders, the fewer resources can be used to make shirts. The opportunity cost of DVD recorders is the quantity of shirts sacrificed by using resources to make DVD recorders not shirts.

Opportunity cost tells us the *relative* cost of producing different goods. The law of comparative advantage states that countries specialize in producing and exporting the goods that they produce at a lower relative cost than other countries. We now develop a model in which differences in relative production costs determine the pattern of international trade. The model demonstrates the **law of comparative advantage**.¹

The **law of comparative advantage** states that countries specialize in producing and exporting the goods that they produce at a lower relative cost than other countries.

Opportunity costs or relative costs may differ in different countries. We begin with a very simple model in which different technology explains the cost difference. Suppose two countries, the US and the UK, produce two goods, DVD recorders and shirts. Labour is the only input and there are constant returns to scale. Table 29.5 shows the assumed production costs. It takes 30 hours of US labour to make a DVD recorder and 5 hours to make a shirt. UK labour is less productive. It takes 60 hours of UK labour to make a DVD recorder and 6 hours to make a shirt.

Table 29.5 Production techniques and costs

	US	UK
Unit labour requirement (hours/output unit)		
DVD recorders	30	60
Shirts	5	6
Wage per hour	\$6	£2
Unit labour cost		
DVD recorders	\$180	£120
Shirts	\$30	£12

For simplicity, assume there is perfect competition. Hence the price of each good equals its marginal cost. With constant returns to scale, marginal costs equal average costs. Thus, prices equal average costs of production. If labour is the only factor of production, average cost is the value of labour input per unit of output – the unit labour cost.

Assume US workers earn \$6 an hour and UK workers £2 an hour. The last two rows of Table 29.5 show unit labour costs of the two goods in each country. With no international trade, each country makes both goods. The unit labour costs are the domestic prices for which the goods are sold. Perfect competition means price equals marginal cost, and constant returns to scale means marginal cost equals average cost.

US unit labour requirements are *absolutely* lower for *both* goods than those in the UK. But US labour is *relatively* more productive in DVD

recorders than in shirts. It takes twice as many labour hours to make a DVD recorder in the UK as it does in the US but only 6/5 times as many hours to make a shirt. These relative productivity differences are the basis for international trade.

Now the countries trade with each other. If each country concentrates on producing the good that it makes *relatively* cheaply, the two countries together make more of *both* goods. Trade leads to a pure gain: extra output to be shared between the two countries. Moreover, the free market provides the right incentives for this beneficial trade to occur.

The countries now trade. Since they use different currencies, a foreign exchange market is set up and an equilibrium exchange rate established. A country's current account must be zero in long-run equilibrium. For simplicity, we ignore foreign debts and assets, and assume that eventually the equilibrium exchange rate adjusts to make the value of imports equal to the value of exports, thus balancing the trade account in the long run.

Table 29.6 shows the unit labour cost and price of smart phones and shirts in different currencies and then shows their price in pounds at three possible exchange rates: \$2.50/£, \$2/£ and \$1.50/£. The domestic prices reflect the unit cost data in Table 29.5. The price in pounds of UK goods is unaffected by the exchange rate, but the UK price of US goods depends on the exchange rate. The more dollars to the pound, the cheaper are both US goods when valued in pounds.

Table 29.6 Costs, prices and the range of equilibrium exchange rates

	Domestic price		Cost in £ at an exchange rate of:					
			\$2.50/£		\$2/£		\$1.50/£	
	Phones	Shirts	Phones	Shirts	Phones	Shirts	Phones	Shirts
US goods	\$180	\$30	£72	£12	£90	£15	£120	£20
UK goods	£120	£12	£120	£12	£120	£12	£120	£12

At \$2.50/£, US phones are cheaper in pounds than UK phones, but the price of UK and US shirts is the same. If the exchange rate exceeds \$2.50/£, even US shirts cost less in pounds. The equilibrium exchange rate cannot lie above \$2.50/£, for then nobody would buy UK goods.² A one-way flow in trade and foreign exchange is not an equilibrium.

Conversely, at \$1.50/£, US shirts are dearer than UK shirts, but UK and US phone prices are the same. If the exchange rate is lower than \$1.50/£, both US goods are dearer than UK goods when valued in the same currency. At \$1/£, US phones cost £180 and US shirts cost £30. At any exchange rate below £1.50/£, there is a one-way flow of trade and foreign exchange, though it is now UK goods everyone wants to buy.

The foreign exchange market is in equilibrium only if the value of UK imports, and hence the demand for dollars with which to purchase them, equals the value of UK exports, and hence the supply of dollars (UK exporters converting revenues back into pounds). The highest possible equilibrium exchange rate is \$2.50/£, the exchange rate at which one UK good (shirts) is still just competitive with US shirts; the lowest possible equilibrium exchange rate is \$1.50/£, the exchange rate at which one US good (phones) is still just competitive with UK goods.

Comparative advantage

Absolute advantage means a country is the lowest-cost producer of that good.

Comparative advantage means the country makes a good relatively more cheaply than it makes other goods.

Table 29.6 shows an intermediate exchange rate, \$2/£. The exact position of the equilibrium exchange rate depends on the demand for DVD recorders and shirts. Regardless of a country's **absolute advantage** in making goods more cheaply, there is always an exchange rate that lets the country make at least one good more cheaply than other countries when all goods are valued in a common currency. At the equilibrium exchange rate, the country has a **comparative advantage** in the production of at least one good which it can then export to pay for its imports.

Although the US has a lower absolute labour requirement for both goods, the relative cost of phones is lower in the US, and the relative cost of shirts higher, than in the UK. In the US, where phones cost \$180 and shirts \$30, the former cost six times as much as the latter. The law of comparative advantages says countries specialize in producing the goods they make relatively cheaply. In the UK – where shirts cost £12 and

phones £120 – the latter cost ten times as much as the former cost. Making phones costs less relative to shirts in the US than in the UK. The *opportunity cost* of phones is lower in the US, which must give up six shirts to make another phone.

Conversely, the opportunity cost of shirts is lower in the UK than in the US. The UK must give up only one-tenth of a phone to make another shirt, compared with one-sixth in the US. The law of comparative advantage says that the UK will specialize in shirts, which have a low opportunity cost for UK producers, and the US will specialize in phones, which have a low opportunity cost for US producers. We discuss the [law of comparative advantage](#) further in Concept 29.1.

The [law of comparative advantage](#) says countries specialize in producing the goods they make relatively cheaply.

Production and trade patterns depend on *comparative* advantage and *relative* costs because the level of the equilibrium exchange rate takes care of differences in absolute advantage. Even if US producers have lower unit labour requirements for both goods, a sufficiently low \$/£ exchange rate makes US goods dear in the UK and UK goods cheap in the US. Beginning from a high \$/£ exchange rate at which no UK goods can compete with US goods, which of the UK goods first becomes competitive as the exchange rate falls? The good in which the UK has a comparative advantage or lower opportunity costs.

The principle of comparative advantage has many applications in everyday life. Suppose two students share a flat. One is faster both at making the dinner *and* at vacuuming the carpet. If tasks are allocated according to absolute advantage, the other student does nothing. The jobs get done fastest if each student does the task at which he is relatively good.

CONCEPT 29.1

Comparative advantage and the gains from trade

The table below summarizes earlier data on unit labour requirements (ULR) in labour hours per unit of output, unit labour cost (ULC) in domestic currency and opportunity cost (OC) in

domestic goods forgone. With lower ULRs, the US has an *absolute advantage* in both goods. One way to calculate *comparative advantage* is to compare ULRs across countries. Relative to the UK, the US needs less labour to produce phones than shirts. The US has a comparative advantage in phones, the UK in shirts.

Alternatively, we can compare opportunity costs, OC. By sacrificing six shirts, the US gets 30 labour hours that make an extra phone. More simply, six shirts cost \$180, the price of one phone. The opportunity cost of a phone is six shirts in the US and ten shirts in the UK. But the opportunity cost of a shirt in the UK (one-tenth of a phone) is less than in the US (one-sixth of a phone). Again, the US has a comparative advantage in phones and the UK in shirts. When there are many factor inputs, this method of calculating comparative advantage is simpler.

The gains from trade

To make 60 shirts, the UK gives up output of 6 phones. To make 6 phones, the US gives up only 36 shirts. Trade and international specialization let the world economy have an extra 24 shirts with no loss of phones. Or if the US makes another ten phones, giving up 60 shirts, the world economy has four more phones with no loss of shirts. These are the *gains from trade*. Only when opportunity costs are the *same* in both countries are there no gains to exploit.

Comparative advantage explains why Europe ought to be able to benefit by opening up trade with Asia, even if Asian producers are very low cost. Europeans can enjoy cheaper goods than before and redeploy resources to more productive alternatives. Of course, this has two implications.

	ULR	UIC	OC
US			
Phones	30	\$180	6 shirts
Shirts	5	\$30	$\frac{1}{6}$ phone
UK			
Phones	60	£120	10 shirts
Shirts	6	£12	$\frac{1}{10}$ phone

First, even if both trading countries benefit on average, there is no guarantee that each and every individual will benefit. For example,

those who specialized in producing the products or services now outcompeted by cheaper imports may be individually worse off. Whether they are actually worse off depends on what the country as a whole does with the gains from trade. It could redistribute some of this gain to the particular individuals who had lost out, thereby ensuring that everybody wins. In practice, it is hard to accomplish complete compensation even if it is attempted.

Second, the gains are largest after redundant resources are redeployed to a better use. At best, this takes time. At worst, it may lead to prolonged inactivity and political unrest.

Governments are often driven to make political commitments – ‘No export of British jobs!’ – but in practice these are both impossible to deliver – the market finds a way – and undesirable to deliver – the whole point of finding a cheaper supplier is to make the switch and allow the purchasing power of consumer incomes to increase. Precluding the switch entails throwing away the potential gains from trade. We study gainers and losers more fully in Case 29.1.

Many goods

The principle of comparative advantage still holds with many goods. Table 29.7 shows a range of commodities. The first two rows show unit labour requirements to make each good in the US and the UK. The third row shows unit labour requirements in the US relative to the UK.

Table 29.7 Unit labour requirements and comparative advantage: many goods (hours of labour input per unit output)

	Computers	Cars	TVs	Textiles	Glass	Shoes
US goods	200	300	50	5	7	15
UK goods	1200	600	90	8	6	10
US/UK relative ULR	1/6	1/2	5/9	5/8	7/6	3/2

Rank the commodities in order. Beginning at the left, the US has the largest comparative advantage in computers, where its relative unit labour requirement is only one-sixth that of the UK. Next is cars, where the US relative unit labour requirement is half that in the UK; then TVs,

textiles, glass and finally shoes. The comparative advantage of the US falls as we move to the right in the table.

Conversely, the UK has the largest comparative advantage in making shoes, the good in which UK producers are most efficient relative to US producers. As we move to the left, the comparative advantage of the UK declines; US producers become increasingly efficient relative to UK producers.

The US has an absolute advantage in producing computers, cars, TVs and textiles. The UK has an absolute advantage in producing glass and shoes. Nevertheless, absolute advantage plays no direct part in the analysis. Comparative advantage is what counts.

The equilibrium exchange rate occurs at some intermediate level that just balances the value of trade between the two countries. Essentially, the level of the exchange rate takes care of the overall level of absolute advantage, leaving comparative advantage to determine trade patterns.

Different capital–labour ratios

Comparative advantage need not depend on technology differences. It may also reflect different factor supplies. Consider the UK and China. The UK has more capital per worker than China. Even though China's vast size may mean that it has absolutely more capital than the UK, the UK has *relatively* more capital than China.

What does this imply about the relative price of hiring labour and capital in the two countries? With more capital per worker, the marginal product of labour is higher in the UK. This makes real wages higher in the UK than in China. Conversely, the number of workers per unit of capital is lower in the UK than in China. The marginal product of capital and the rental of capital will tend to be lower in the UK, where machinery is relatively plentiful, than in China, where machinery is relatively scarce. Because the UK is endowed with more capital relative to labour than China, the cost of using labour relative to capital is higher in the UK than in China.

Relative costs of using inputs affect the relative price of the goods they produce. Goods made by labour-intensive methods cost relatively more to make in the UK than in China. Suppose car production is capital intensive with sophisticated assembly lines, but textile production is

labour intensive with detailed tasks best done by hand. The price of cars relative to textiles is lower in the UK than in China.

Hence, a *relatively* abundant supply or endowment of one factor of production tends to make the cost of renting that factor relatively cheap. Goods that use that factor relatively intensively are thus relatively cheap. In these goods the country has a comparative advantage. Thus the UK, relatively generously supplied with capital relative to labour, exports capital-intensive cars to China. China, relatively well endowed with labour, should export labour-intensive textiles to the UK. Differences in relative factor supply are an important explanation for comparative advantage and the pattern of international trade.

Figure 29.1 supports this analysis. It emphasizes skills, or human capital, rather than physical capital, although the two are usually correlated. Countries with scarce land but abundant skills have high shares of manufactures in their exports. Countries with lots of land but few skills typically export raw materials. The figure also shows regional averages. Africa lies at one end, the industrial countries at the other.

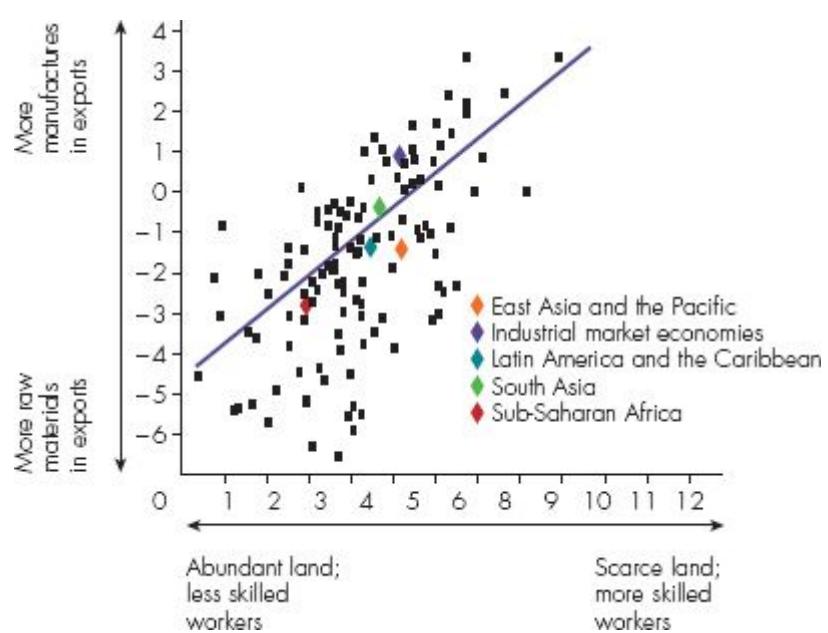


Figure 29.1 Comparative advantage and export composition (125 countries and regional averages)

Source: World Bank, World Development Report, 1995.

We now have two explanations for comparative advantage or international differences in relative production costs. The first is the Ricardian explanation – international differences in technology that cause differences in relative physical productivity and relative unit labour requirements. Second, even if countries have access to the same technology, the domestic relative price of goods may differ across countries because the relative cost of renting factor inputs differs across countries. Where a factor is in relatively abundant supply, goods that use that factor relatively intensively are likely to be relatively cheaper than in other countries.³

MATHS 29.1

THE GAINS FROM TRADE

In Chapter 1, we introduced the production possibility frontier, which typically has a concave shape because of diminishing marginal returns to each activity. For simplicity, we ignore diminishing returns, implying that production possibility frontiers are straight lines.

There are two goods, X and Y, and two countries, A and B. Country A is endowed with 120 units of the variable resource (think of it as labour), which can be allocated to either industry. Country A's production possibility frontier is given by Both versions of equation (1) show the maximum amount of good Y that country A can produce, given the level of good X being produced. For country B, which is much smaller and has only 40 units of resources to allocate, its equations happen to be:

$$120 = Y_A + 2X_A \quad Y_A = 120 - 2X_A \quad (1)$$

$$40 = Y_B + X_B \quad Y_B = 40 - X_B \quad (2)$$

Suppose in each country, consumers like to consume the same quantity of X and Y. From equations (1) and (2), this implies:

$$Y_A = X_A = 40 \quad Y_B = X_B = 20 \quad (3)$$

Now suppose international trade is possible. Which country has a comparative advantage in which good? In country A, sacrificing one unit of X leads to an extra 2 units of Y. In country B, sacrificing one unit of X allows an extra 1 unit of Y. Hence, country A is

relatively better at making Y and country B is relatively better at making X.

Suppose the small country specializes completely in making the good in which it has a comparative advantage. Thus,

$$Y_B = 0 \quad X_B = 40 \quad (4)$$

Country A will largely make good Y but remember that in this (strange) world, consumers demand equal quantities of both goods. Any extra of only one good is completely worthless. So country A may have to make some of the other good as well.

In total,

$$120 = Y_A + 2X_A \quad \text{and} \quad 80 = 2X_B$$

So adding these together, $200 = Y + 2X$, where Y and X denote total world production. Since consumers want equal quantities of each, the best outcome is where

$$X = Y = 66.7 \quad (5)$$

This implies that country A makes 66.67 of good Y and $(66.67 - 40 = 26.67)$ of good X, which is consistent with equation (1).

Before international trade, country A made 40 of each good and country B made 20 of each good, so world output was 60 of each good. Equation (5) confirms the gains from trade. By focusing on the good in which they have a comparative advantage, world output, and potential consumption, of both goods increases.

29.3 Intra-industry trade

Intra-industry trade is two-way trade in goods made within the same industry.

Different countries have a comparative advantage in different goods and specialize in producing these goods for the world economy. This explains why the UK exports cars to China but imports textiles from

China. It does not explain why the UK exports cars (Jaguar, Aston Martin, Mini) to Germany while simultaneously importing cars (Mercedes, BMW, VW, Audi) from Germany.

A Jaguar is not exactly the same commodity as a Mercedes, nor is Carlsberg exactly the same as Stella. Cars and beer are industries each making a range of different, but highly substitutable, products which enjoy brand allegiance.

Intra-industry trade reflects three factors. First, consumers like a wide choice of brands. They do not want exactly the same car as everyone else. Second, there are important economies of scale. Instead of each country making small quantities of each brand in each industry, it makes sense for the UK to make Jaguars, Germany to make Mercedes and Sweden to make Volvos and then to swap them around in international trade. Third, the tendency to specialize in a particular brand, to which the demand for diversity and the possibility of scale economies give rise, is limited by transport costs. Intra-industry trade between Germany and France is larger than intra-industry trade between Germany and Japan.

To measure the importance of intra-industry trade, we define an index as zero when trade in a particular commodity is entirely one-way: a country either exports or imports the good, but not both.

At the opposite extreme, the index equals 1 when there is a complete two-way trade in a commodity: a country imports as much of the commodity as it exports. Figure 29.2 shows the index for trade by a typical developed economy.

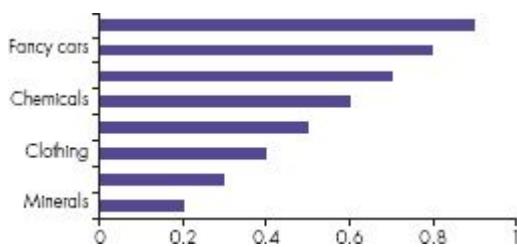


Figure 29.2 Developed economy trade patterns
Index of two-way trade

0 = all 1-way trade, comparative advantage

1 = all 2-way trade, scale economies and diversity

At one extreme, in clothing there is little two-way trade. The US imports clothing but exports very little. This trade obeys the principle of comparative advantage. Similarly, Africa largely exports minerals but does not import them. At the other extreme is banking. Here, trade is almost completely two-way. There are French banks in London and British banks in Paris. As a general principle, the more commodities are undifferentiated goods (fuel, steel, oil), the more trade patterns reflect comparative advantage based on relative resource abundance. As we move towards finished manufactures, product differentiation becomes dominant and comparative advantage loses some of its overriding role. Intra-industry trade is more significant in cars, banking and indeed universities – where French students study in Germany and German students study in France.

The more closely markets are integrated, and the lower the obstacles to trade – in terms of both distance and tariffs – the more intra-industry trade we expect. Japan, geographically isolated from the US and Europe, engages in quite a lot of one-way trade with these markets. In contrast, the EU has a more diversified resource endowment and a more integrated market, in which distance, information barriers and tariffs are now small. Intra-industry trade is extensive. The gain from trade is not the exploitation of differences in relative prices across countries but greater diversity and the lower unit costs that scale economies allow.

CASE 1.1

Historical gainers and losers from trade

Countries trade because they have a comparative advantage (based either on a relative advantage in technology or on relative factor abundance) or because different countries specialize in making different brands when scale economies exist. Either way, countries buy goods more cheaply than they could have done without international trade. Although trade is beneficial in the aggregate, this is no guarantee that trade makes *everyone* better off. Current concerns about globalization arise because there are losers too. Here are some historical examples of the conflicts to which international trade gave rise.

Refrigeration

At the end of the nineteenth century, the invention of refrigeration enabled Argentina to supply frozen meat to the world market. Argentina's meat exports, non-existent in 1900, rose to 400 000 tonnes a year by 1913. The US, with beef exports of 150 000 tonnes in 1900, had virtually stopped exporting beef by 1913.

Who gained and who lost in this early example of globalization? Argentina's economy was transformed. Owners of cattle and land gained; other land users lost out because higher demand bid up land rents. Argentinian consumers found their steaks becoming dearer as meat was shipped abroad. Argentina's GNP rose a lot, but the gain from trade was not equally distributed. Some people in Argentina were worse off. Similarly, in Europe and the US, cheaper beef made consumers better off. But beef producers lost out because beef prices fell. Landowners experienced an overall drop in the demand for land, and had to reallocate it to other, less profitable, uses.

Refrigeration opened up the world to Argentinian beef. As a whole, the world economy gained. In principle, the gainers could have compensated the losers and still had something left over. In practice, gainers rarely compensate losers. So some people lost out. In this example the biggest losers were beef producers elsewhere in the world and other users of land in Argentina.

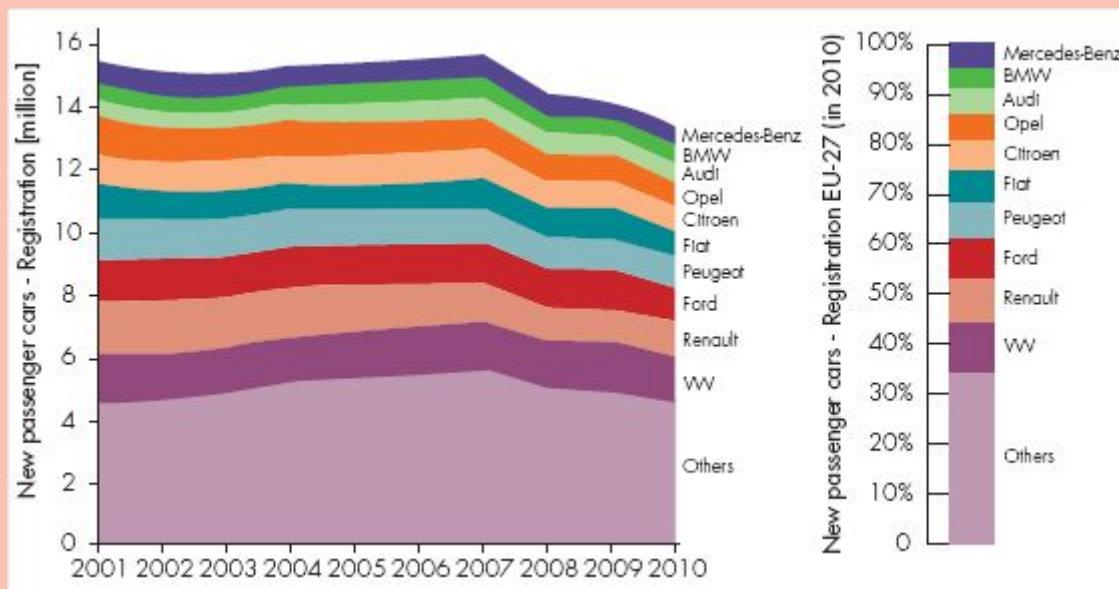
The modern car industry

A second example is the car industry. As recently as 1971, imports of cars were only 15 per cent of the domestic UK market, while 35 per cent of UK car output was exported. The UK was a net exporter of cars. After 1971, UK car makers lost market share to foreign imports. Imports then exceeded 60 per cent of the UK market, but this was not the end of the story. Many foreign car producers decided to build car factories in the UK. UK exports then recovered in the 1990s as Nissan, Honda and Toyota built cars in the UK for export to the EU market.

UK car buyers and foreign car exporters benefited from the rise in UK imports of cheaper foreign cars. British consumers no longer bought unreliable (and relatively expensive) Austin Allegros, Morris Marinas and Triumph Dolomites. They enjoyed VW Golfs, Nissan Micras and Toyota Corollas instead. The original UK car producers had a very tough time, and lost out from greater international trade. The Mini was bought by BMW in 1994; Aston

Martin was sold to Ford in 1997, Bentley to Volkswagen in 1998, Rover to Nanjing Automobile of China in 2005 and Jaguar Land Rover to the Indian Tata Group in 2008. In a global automotive market, global brands need large investments in new models and factories. The UK was too small to sustain the range of car producers that it had in the 1970s.

Today, France has the same dilemma: Renault, Peugeot and Citroen struggle in a global car business with too many small producers fully to exploit scale economies. In late 2012, the Committee of French Automobile Producers reported that sales of Citroen had fallen by 26 per cent in a year, and sales of state-owned Renault had fallen by 33 per cent. As elsewhere in Europe, French buyers under financial pressure were turning to Japanese Toyota, Nissan and Honda, but also to cheaper Koreans, Hyundai and Kia, as shown in the figure below. Every time a major French car manufacturer proposes job reductions or the closure of a factory in order to rationalize production, French trade unions exert considerable power of protest. French governments usually lean on the car company to back down. French car workers keep their jobs a little longer, and French consumers continue to lose out.



Source: European Vehicle Market Statistics Pocketbook 2012, by Peter Mock, 2012, © 2012 International Council on Clean Transportation.

Restricting car imports helps local car producers but raises car prices for local car buyers. Should the government please producers or consumers? More generally, is restricting imports a good idea?

In the next section we analyse the costs and benefits of tariffs or other types of trade restriction. In so doing, we move from *positive economics*, why trade exists and what form it takes, to *normative economics*, what trade policy is desirable.

The steel industry

As the workshop of the world in the nineteenth century, the UK pioneered industries such as steelmaking. An early technological lead, and convenient deposits of raw materials, provided an initial comparative advantage. As the industry prospered, the UK acquired skills – human capital – that further consolidated its comparative advantage.

Gradually, of course, other countries learned the technology and trained their workers. Britain lost its unique lead, and was then overhauled by countries that had more easily mined deposits of coal and iron, cheaper wages or less crowded locations in which low-cost modern factories were more easily assembled.

Initially, this additional global capacity was built in Europe and the US, then in Japan. By the 1970s many global sources of cheaper steel were making European steel less competitive. European governments were drawn into state subsidies and an exit game among themselves – if some countries sharply contracted their steel industry, world prices would rise and other European steel producers might survive for a while longer. But no country wanted to take the pain of cutbacks that would largely benefit its neighbours.

Within the UK, British Steel was nationalized in 1967 and privatized again in 1988, seeking a global niche as a producer of high-quality and high-priced steel. Unable to succeed on its own, it then merged with a Dutch company in 1999 to form Corus, which in turn was taken over in 2007 by Tata Steel, part of the Indian conglomerate Tata Sons. In 2010 Tata Steel announced plans to mothball indefinitely its Middlesbrough steel plant, an iconic symbol of heavy industry in the north-east of England. However, in 2012 a group of Thai investors announced plans to reopen steel mills in the area.

Even so, like the UK coal industry, the UK steel industry is well on the way to disappearing, as comparative advantage evolves over time. If evolution did not occur, the UK would still be a land of handloom weavers, indeed of hunter gatherers. Against these costs of global competition we have to weigh: (a) the consumer benefits of obtaining similar or better products more cheaply through imports, and (b) the productivity gains of diverting the labour force out of old industries into the industries of the future.

29.4 The economics of tariffs

Trade policy affects international trade through taxes or subsidies, or by direct restrictions on imports and exports.
A **tariff** is a tax on imports.

We now turn from the determinants of international trade to international **trade policy**.

The most common type of trade restriction is a **tariff** or import duty.

If t is the tariff rate as a decimal fraction, the domestic price of an imported good is $(1 + t)$ multiplied by its world price. By raising the domestic price of imports, a tariff helps domestic producers but hurts domestic consumers.

The free trade equilibrium

Figure 29.3 shows the domestic market for cars. The UK faces a given world price of cars, say £10 000 a car, shown by the solid horizontal line. Schedules DD and SS show the demand for cars by UK consumers and the supply of cars by UK producers. For simplicity, we assume that domestic and foreign cars are perfect substitutes. Consumers buy the cheaper one.

At a price of £10 000, UK consumers wish to buy Q_d cars, at point G on their demand curve. Domestic firms wish to make Q_s cars at this price. The gap between domestic supply Q_s and domestic demand Q_d is imported.

Equilibrium with a tariff

Suppose that the government levies a 20 per cent tariff on imported cars. Car importers charge £12 000 to cover their costs, inclusive of the tariff. The broken horizontal line at this price shows that importers will sell any number of cars in the domestic market at a price of £12 000. The tariff raises the domestic tariff-inclusive price above the world price.

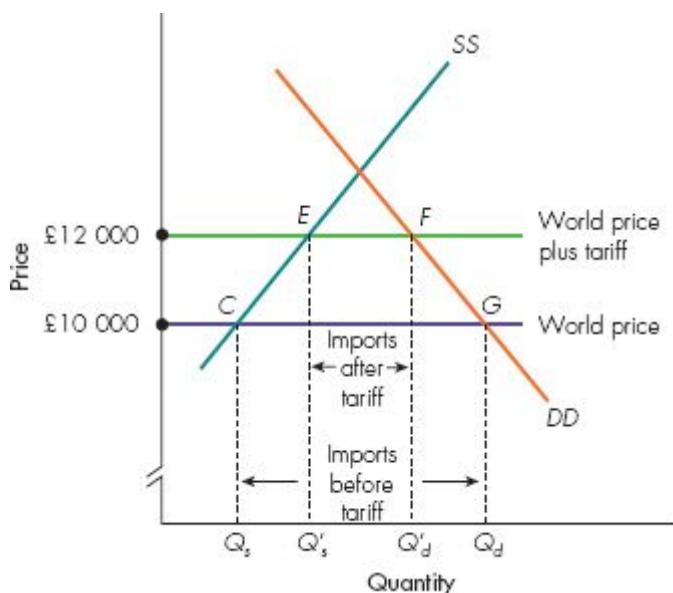


Figure 29.3 The effects of a tariff

DD and SS show the domestic demand and supply for cars. In the absence of a tariff, consumers can import cars at a price of £10 000. In free trade equilibrium, domestic producers produce at C and domestic consumers consume at G. The quantity of imported cars is CG. Qd is the total quantity demanded. Domestic production Qs is supplemented by imports ($Q_d - Q_s$). A 20 per cent tariff raises the domestic price of imports to £12 000. Domestic output is now at E and consumers consume at F. Imports fall from CG to EF.

By raising domestic car prices, the tariff boosts domestic car production from Q_s to Q'_s . The tariff protects domestic producers by raising the domestic price at which imports become competitive. In moving up the supply curve from C to E, domestic producers with marginal costs between £10 000 and £12 000 can now survive at the higher domestic price of cars.

The higher price also moves consumers up their demand curve from G to F. The quantity of cars demanded falls from Q_d to Q'_d . For consumers,

the tariff is like a tax. Cars cost more.

Figure 29.3 shows the combined effect of higher domestic production but lower domestic consumption. Imports fall because domestic production rises *and* because domestic consumption falls. The more elastic are these supply and demand schedules, the more a given tariff reduces imports. If both schedules are very steep, the quantity of imports hardly changes.

Costs and benefits of a tariff

Figure 29.4 shows the costs and benefits of imposing a tariff. We distinguish *net costs to society* from *transfers* between one part of the economy and another.

After the tariff is imposed, consumers buy the quantity Q'_d . Since the consumer price rises by £2000, consumers spend ($\text{£}2000 \times Q'_d$) more than before to buy the quantity Q'_d . Who gets these extra payments – the area *LFHJ* in Figure 29.4?

Some of the extra consumer payments go to the government. Tariff revenue is the rectangle *EFHI*, a tariff of £2000 per imported car multiplied by $(Q_d - Q'_s)$ the number of imported cars. This transfer *EFHI* from consumers to government is *not* a net cost to society. The government could use the tariff revenue to reduce income tax rates.

Higher consumer payments also go in part to firms as extra profits. Firms get a higher domestic price for their output. The supply curve shows how much firms need to cover the extra cost of making Q'_s not Q_s . Hence the area *ECJL* shows extra profits; namely, the extra revenue from higher prices not required to cover extra production costs. *ECJL* is a transfer from consumers to the pure profit of firms. It is *not* a net cost to society.

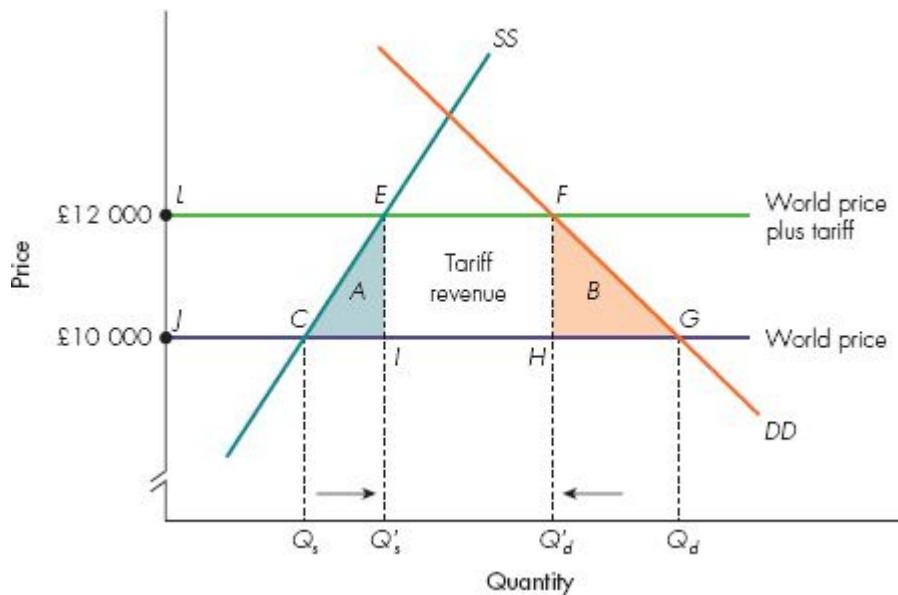


Figure 29.4 The welfare costs of a tariff

The tariff leads to both transfers and to net social losses. The tariff raises the domestic price from £10 000 to £12 000. LFHJ shows extra consumer payments of the Q'_d cars they now buy. But EFHI is a transfer to the government and ECJL is a transfer to extra profits of producers. Areas A and B are pure waste and net social losses. Triangle A is the extra that society spends by producing cars domestically instead of importing them at the world price. Triangle B is the excess of consumer benefits over social marginal cost that society sacrifices by reducing its consumption of cars from Q_d to Q'_d .

The shaded area *A* is part of the area *LFHJ* showing extra consumer payments, but not government revenue or extra profits for firms. It is a net cost to society: the cost of supporting inefficient domestic firms. The supply curve *SS* shows the marginal cost of making the last car in the home economy. But society *could* import cars from the rest of the world in unlimited quantities at the world price, which is the true marginal cost of cars to the domestic economy. The triangle *A* shows the resources that society wastes by making the quantity ($Q'_s - Q_s$) at home when it could import at a lower cost. Resources drawn into domestic car production could be better used elsewhere in the economy.

Triangle *B* is a second net loss to society. If the tariff is scrapped and free trade restored, the quantity of cars demanded rises to Q_d . The triangle *B* shows the excess of consumer benefits, measured by the height of the demand curve (how much consumers will pay for the last unit demanded) over the marginal costs of expanding from $Q'_d - Q_d$, the

world price at which imports can be purchased. Conversely, by imposing the tariff, society incurs a net loss shown by triangle *B*. It is the net benefit society gives up when fewer cars are bought by consumers.

To sum up, starting from free trade equilibrium and then imposing a tariff, the higher domestic price leads to transfers and to pure waste. Money is transferred from consumers to the government and to producers. The net social cost of these transfers is zero.

A tariff also involves pure waste. Society can always import cars at the world price. By raising domestic prices, a tariff leads to domestic overproduction and domestic under-consumption of the good. Triangles *A* and *B* measure this waste. Society does better to use fewer resources in the car industry and to transfer these resources to an export industry that could earn enough foreign exchange to import cars at the cheaper world price. This is the *case for free trade*.

29.5

Good and bad arguments for tariffs

Table 29.8 lists some arguments for tariffs. We group them under several headings. The *first-best* argument is a case where a tariff is *the* best way to achieve a given objective. *Second-best* arguments are cases where a tariff is better than nothing, but where another policy is better still if it can be implemented. Non-arguments are fallacious.

Table 29.8 Arguments for tariffs

Type	Example
First-best	Foreign trade monopoly
Second-best	Way of life, anti-luxury, infant industry, defence, revenue
Strategic	Games against foreigners
Non-argument	Cheap foreign labour

The optimal tariff: the first-best argument for tariffs

When imports affect the world price, the **optimal tariff** reduces imports to the level at which social marginal cost equals social marginal benefit.

In presenting the case for free trade, we were careful to assume that the domestic economy can import as many cars as it wished without bidding up the world price of cars. For a small economy this is a reasonable assumption. However, imports by a large country may be large relative to the world market, and bid up the world price of those commodities.

In this case, the world price of the last unit imported is *lower* than the true cost of the last unit to the importing economy. In demanding another unit of imports, the economy raises the price it has to pay on the quantity already being imported. In a free trade world without tariffs, each individual thinks only about the price she pays. No single individual bids up the price, but collectively individuals bid up the price of imports.

Under free trade, each person buys imports until the benefit to that individual equals the world price she pays. Since the collective cost of the last import exceeds its world price, the social cost of that import exceeds its benefit. There are too many imports. Society gains by restricting imports until the benefit of the last import equals its social cost. The optimal tariff is a straightforward application of the principles of efficient resource allocation discussed in Part Three.

A small country's imports have no effect on the world price of its imports. The marginal social cost of imports equals the world price. Then, and only then, is the optimal tariff zero. Free trade is then first-best.

Second-best arguments for tariffs

The **principle of targeting** says that the most efficient way to attain a given objective is to use a policy influencing that activity directly.

We now introduce the **principle of targeting**.

The optimal tariff is a first-best application of the principle of targeting precisely because the source of the problem is a divergence between social and private marginal costs in trade itself. A tariff on trade is the most efficient solution. The arguments for tariffs that we now examine are all second-best arguments because the original source of the problem does not directly lie in trade. The principle of targeting assures us that there are ways to solve these problems at a lower net social cost.

Way of life

Suppose society wishes to help inefficient farmers or craft industries. It believes that the old way of life, or sense of community, should be preserved. It levies tariffs to protect such groups from foreign competition.

There is a cheaper way to attain this objective. A tariff helps domestic producers but hurts domestic consumers through higher prices. A production subsidy would still keep farmers in business and, by tackling the problem directly, would avoid hurting consumers. In terms of Figure 29.4, triangle *A* shows the net social cost of subsidizing domestic producers so they can produce Q_s^* rather than Q_s . But a tariff, the second-best solution, also involves the social cost given by the triangle *B*.

Suppressing luxuries

Some poor countries believe it is wrong to allow their few rich citizens to buy Rolls-Royces or luxury yachts when society needs their resources to stop people starving. A tariff on imports of luxuries reduces their consumption but, by raising the domestic price, may also provide an incentive for domestic producers to use scarce resources to produce them. A consumption tax tackles the problem directly, and is more efficient.

Infant industries

A common argument for a tariff is that it allows infant industries to get started. Suppose there is *learning by doing*. Only by actually being in business do firms learn how to reduce costs and become as efficient as foreign competitors. A tariff provides protection to infant industries until they master the business and can compete on equal terms with more experienced foreign suppliers.

Society should invest in new industries only if they are socially profitable in the long run. The long-run benefits must outweigh the initial losses during the period when the infant industry is producing at a higher cost than the goods could have been obtained at through imports. But in the absence of any divergence between private and social costs or

benefits, an industry will be socially profitable only if it is privately profitable.

If the industry is such a good idea in the long run, society should begin by asking why private firms cannot borrow the money to see them through the early period when they are losing out to more efficient foreign firms. If private lenders are not prepared to risk their money, society should ask whether the industry is such a good idea after all. And if the industry does make sense but there is a problem in the market for lending, the principle of targeting says that the government should intervene by lending money to private firms.

Failing this, a production subsidy during the initial years is still better than a tariff, which also penalizes consumers. The worst outcome is a *permanent* tariff, which allows the industry to remain sheltered and less efficient than its foreign competitors long after the benefits of learning by doing are supposed to have been achieved.

Defence

Some countries believe that, in case there is a war, it is important to preserve domestic industries that produce food or jet fighters. Again, a production subsidy, not an import tariff, is the most efficient way to meet this objective.

Revenue

In the eighteenth century, most government revenue came from tariffs. Administratively, it was the simplest tax to collect. Today, this is still true in some developing countries. But in modern economies with sophisticated systems of accounting and administration, the administrative costs of raising revenue through tariffs are not lower than the costs of raising revenue through income taxes or taxes on expenditure. The balance of tax collection should be determined chiefly by the extent to which taxes induce distortions, inefficiency and waste, and the extent to which they bring about the distribution of income and wealth desired by the government. The need to raise revenue is not a justification for tariffs themselves.

Strategic trade policy

In Part Two we argued that game theory is useful in analysing strategic conflict between oligopolists. In international trade, strategic rivalry may exist either between the giant firms or –national champions— of different countries, or between governments acting on their behalf. Strategic international competition may justify industrial policy. Initial subsidy of European aircraft producer Airbus Industrie was a pre-commitment designed to deter Boeing from trying to force Airbus out of the industry.

Similar considerations arise in trade policy. Levying a tariff on imports, thereby protecting domestic producers, may deter foreigners from attempting a price war to force the domestic producers out of the industry, and may prevent foreign producers from entering the industry.

This sounds like a very general and robust argument for tariffs, but it should be viewed with considerable caution. If it is attractive for one country to impose tariffs for this purpose, it may be as attractive for foreigners to retaliate with tariffs of their own. We then reach an equilibrium in which little trade takes place, domestic giants have huge monopoly power since they no longer face effective competition from foreigners, and all countries suffer.

In fact, this game has the structure of that of the Prisoner's Dilemma, which we introduced in Part Two. All countries may be led to impose tariffs even if all would be better off if tariffs were abolished. There may be a role for international co-operation to agree on, and subsequently enforce, low tariffs. We take up this theme shortly.

Dumping

Dumping occurs when foreign producers sell at prices below their marginal production cost, either by making losses or with the assistance of government subsidies.

The preceding discussion relates to tariffs, but also applies to trade subsidies. Domestic producers say this is unfair and demand a tariff to protect them from foreign competition. If we knew foreign suppliers would supply cheap goods indefinitely, we should say thank you, close down our more expensive industry and put our resources to work elsewhere. To this extent, dumping is a non-argument for a tariff.

However, foreign producers may be engaged in predatory pricing meant to drive our producers out of the industry. Once the foreigners achieve monopoly power in world markets, they raise prices and make big profits. If so, it may be wise for our government to resist. Even so, a production subsidy is the efficient way to insulate our producers from this threat. A tariff has the undesirable side effect of distorting consumer prices.

Non-arguments for tariffs

Cheap foreign labour

Home producers frequently argue that tariffs are needed to protect them from cheap foreign labour. Yet the whole point of trade is to exploit international differences in the relative prices of different goods. If the domestic economy is relatively well endowed with capital, it benefits from trade precisely because its exports of capital-intensive goods allow it to purchase *more* labour-intensive goods from abroad than would have been obtained by diverting domestic resources to production of labour-intensive goods.

As technology and relative factor endowments change over time, countries—comparative advantage alters. In the nineteenth century Britain exported Lancashire textiles all over the world. But textile production is relatively labour intensive. Once the countries of Southeast Asia acquired the technology, it was inevitable that their relatively abundant labour endowment would give them a comparative advantage in producing textiles.

New technology frequently gives a country a temporary comparative advantage in particular products. As time elapses, other countries acquire the technology, and relative factor endowments and relative factor costs become more important determinants of comparative advantage. Inevitably, the producers who have lost their comparative advantage start complaining about competition from imports using cheap foreign labour.

In the long run, the country as a whole will benefit by facing facts, recognizing that its comparative advantage has changed and transferring production to the industries in which it now has a comparative advantage. Our analysis of comparative advantage promises us that there *must* be an industry in which each country has a comparative advantage.

In the long run, trying to use tariffs to prop up industries that have lost their comparative advantage is futile and costly.

Of course, in the short run the adjustment may be painful. Workers lose their jobs and must start afresh in industries in which they do not have years of experience and acquired skills. But the principle of targeting tells us that, if society wants to smooth this transition, some kind of retraining or relocation subsidy is more efficient than a tariff.

Why do we have tariffs?

Aside from the optimal tariff argument, there is little to be said in favour of tariffs. Economists have argued against them for over a century. Why are tariffs still so popular?

ACTIVITY 29.1

The EU Single Market

The European Community was founded in 1957 as a free trade area – abolishing tariffs and quotas on trade between member states. Over the next 50 years, the original six – West Germany, France, Italy, Netherlands, Belgium and Luxembourg – were joined by Denmark, Ireland and the UK in the 1970s, by Spain, Portugal and Greece in the 1980s, and by Austria, Finland and Sweden in the 1990s. The European Community (EC) became the European Union (EU). In 2004 the EU admitted the Baltic republics (Estonia, Latvia, Lithuania), the countries of central Europe (Hungary, Poland, Czech Republic, Slovakia, Slovenia) and the Mediterranean islands of Malta and the Greek part of Cyprus. Bulgaria and Romania were admitted in 2007.

EU enlargement was not initially accompanied by any change in its fundamental structure. Member states still set national policies. Harmonization was thwarted for two reasons. First, since each country did things differently, it was hard to find a single set of regulations for all member states. Second, it was political dynamite. No country wanted to adopt the policies of others. What unlocked the process was agreement that firms licensed by the rules of their home country could then operate in other member states.

By 1992 a single EU market among member states was established, involving: (a) free capital flows between members; (b) removal of all non-tariff barriers to trade in the EU (different trademarks, patent laws and safety standards that made it hard for imports to compete with domestic goods even when tariffs were zero; (c) ending of the bias in public sector purchasing to favour domestic producers; (d) removal of frontier controls (delays); and (e) progress in harmonizing tax rates. Tariffs and quotaNon-tariff barriers are different national regulations or practices that prevent free movement of goods, services and factors between countries.s are not the only form of trade protection.

By removing **non-tariff barriers**, the Single Market aimed to allow countries to exploit their comparative advantage more fully.

A second inefficiency in small and segmented national markets is that firms cannot fully exploit economies of scale. As barriers came down, firms got larger and costs fell.

Non-tariff barriers are different national regulations or practices that prevent free movement of goods, services and factors between countries.

The Single Market intensified competition in two ways. First, competition between forms of regulation led on average to lower levels of regulation. For many continental European countries, the Single Market led to substantial deregulation from initial levels that had been very high. Second, a larger market enabled large firms to enjoy scale economies *without* the high market share and potential monopoly power that this would have meant in small, segmented economies.

Quantifying the gains

How large were the gains in practice? In 2002 the European Commission estimated that during 1992–2002 the Single Market had raised members— GDP by 1.8 per cent above what it otherwise would have been (a gain of €5700 per household, which is quite substantial), and had also raised employment by 1.46 per cent.

In general, small countries gained more than large countries, but gains also reflected the pattern of trade. The largest gains came as

the most protected activities were opened up. Not only was the Single Market good for the EU, it also turned out to boost trade with the outside world. Fears of fortress Europe were unfounded.

Benefits of the Single Market

The Single Market is a *wider market* comprising nearly 380 million consumers and making up almost 40 percent of world trade. Such a huge market gives consumers greater choice.

The greater *competition and liberalization* that the Single Market helped to bring about led to *lower prices*. Cheap airlines such as easyJet would not have been possible without the Single Market. Airlines can now fly where they want, without national restrictions. BA became the second-largest domestic airline in France.

The Single Market provides for better *consumer protection*; for example, the Toy Directive means that all toys sold in the EU must be safe for children. Another example is the Motor Insurance Directives, making it easier for those involved in motor accidents in other member states to make an insurance claim when returning home.

The Single Market principle of mutual recognition of standards meant a significant reduction in export bureaucracy. The Single Market is in effect a domestic market for European business.

Each member state's citizens have the *right to work, study or retire* in all the other member states.

Questions

- (a) Gains from trade arise either from exploiting comparative advantage and engaging in more one-way trade, or from achieving greater scale economies and diversity by engaging in more two-way trade in the same commodities or services. Did the largest gains from the EU Single Market programme arise from one-way or two-way trade in the product? Why do you think this?
- (b) If trade also leads to greater competition, which of the above two channels are gains from trade arising through?
- (c) Could the removal of non-tariff barriers lead to greater one-way trade? Could Figure 29.4 be amended to display this?

To check your answers to these questions, go to page 682.

Concentrated benefits, diffuse costs

A tariff on a particular commodity helps a particular industry. It is easy for firms and workers in an industry to organize effective political pressure, for they can all agree that this single issue is central to their livelihood, at least in the short run. But if the tariff is imposed, the cost in higher consumer prices is borne by a much larger and more diverse group of people whom it is much harder to organize politically. Hence, politicians heed the vociferous, well-organized group lobbying *for* tariffs, especially if they are geographically concentrated in an area where, by voting together, they have a significant effect on the outcome of the next election.

Tariffs vs subsidies

Why does government assistance often take the form of tariffs rather than production subsidies, which are frequently more appropriate? If domestic industry is suffering from imports of Chinese goods, the solution seems to be to hurt China. Moreover, if the government instead utilized a production subsidy, it would have to raise other taxes to finance this subsidy. A tariff is often politically easier because it seems to augment government revenues (raising hopes of an income tax cut), whereas a subsidy seems to deplete government revenues (raising fears of higher tax rates). A tariff hits consumers directly by raising the domestic price of the good, but the government may be able to invoke impersonal –market forces—. Tariffs cause the government less political hassle.

29.6 Tariff levels: not so bad?

In the nineteenth century world trade grew rapidly, in part because the leading country, the UK, had a vigorous policy of free trade. US tariffs were about 50 per cent, although they had fallen to around 30 per cent by the early 1920s. As the industrial economies were hit by the Great Depression of the late 1920s and 1930s, there was pressure to protect domestic jobs by keeping out imports. Tariffs in the US returned to around 50 per cent, and the UK abandoned the policy of free trade that

had been pursued for nearly a century. Table 29.1 showed that the combination of world recession and tariff wars led to a slump in the volume of world trade, further exacerbated by the Second World War.

The World Trade Organization

After the war, there was a collective determination to see world trade restored. Bodies such as the International Monetary Fund and the World Bank were set up and many countries signed the General Agreement on Tariffs and Trade (GATT) – a commitment to reduce tariffs successively and dismantle trade restrictions.

Under successive rounds of GATT, tariffs fell steadily. By 1960 US tariffs were only one-fifth of their level at the outbreak of the Second World War. In the UK the system of wartime quotas on imports had been dismantled by the mid-1950s, after which tariffs were reduced by nearly half in the ensuing 25 years. Europe as a whole moved towards an enlarged EU in which tariffs between member countries were abolished.

The GATT Secretariat, now called the World Trade Organization (WTO), began the latest round of negotiations – the Doha Development Round in Qatar in 2001. Chinese membership of the WTO has now been agreed. The WTO is increasingly associated with pressure not only to dismantle substantial protection that severely reduces efficiency but also to extend trade liberalization to more and more countries. Tariff levels throughout the world are as low as they have ever been. Trade liberalization has been an engine of growth. World trade has seen six decades of rapid growth.

Nevertheless, the Doha Development Round is not yet a success. As emerging market economies become more successful, they become more powerful in trade negotiations. The most significant differences are between developed nations – led by the US, the EU and Japan – and major emerging market countries led by India, Brazil, China and South Africa. There is also considerable contention against and between the EU and the US over their maintenance of agricultural subsidies, which are seen to operate effectively as trade barriers.

Nor will trade negotiations solve other problems. Fears about globalization often relate to other distortions that more trade then exacerbates. If environmental protection and corporate accountability are weak, globalization may allow environmental exploitation.

The principle of targeting also tells us that the best solution may not be to hinder trade but rather to attack the problems at source. For example, strengthening environmental protection may be a more effective response than perpetuating trade restrictions.

29.7 Other trade policies

Quotas are restrictions on the maximum quantity of imports.

Export subsidies are government assistance to domestic firms in competing with foreign firms.

Tariffs are not the only form of trade policy. We now examine three other policy instruments: **quotas**, non-tariff barriers and **export subsidies**.

Quotas restrict the *quantity* of imports, but also affect domestic prices of the restricted goods. With a lower supply, the equilibrium price is higher than under free trade. Quotas are rather like tariffs. The higher domestic price allows inefficient domestic producers to produce a higher output than under free trade. Quotas lead to social waste for the same reasons as tariffs.

Because quotas raise the domestic price of the restricted good, the lucky foreign suppliers who succeed in getting some of their goods sold make large profits on these sales. In terms of Figure 29.4, the rectangle *EFHI*, which would have accrued to the government as revenue from a tariff, now accrues to foreign suppliers or domestic importers as extra profit. It is the difference between domestic and world prices on the goods that are imported, multiplied by the quantity of imports allowed.

If these profits accrue to foreigners they are a social cost of quotas over and above the cost of an equivalent tariff. However, the government could always auction licences to import and recoup this revenue. Private importers or foreign suppliers would bid up to this amount to acquire an import licence.

Non-tariff barriers include delaying imports at the frontier, a home-goods bias in government procurement and contracts that specify standards with which domestic producers are familiar but foreign producers are not.

So far, we have looked at restrictions on imports. Countries also use trade policy to boost exports. This can vary from outright subsidy to cheap credit or exemption from certain domestic taxes. Figure 29.5 shows the economics of an export subsidy. Suppose the world price of a good is £5000. Under free trade, domestic consumers buy a quantity Q_d at point G on their demand curve, producers make a quantity Q_s at point E on their supply curve and a quantity GE is exported.

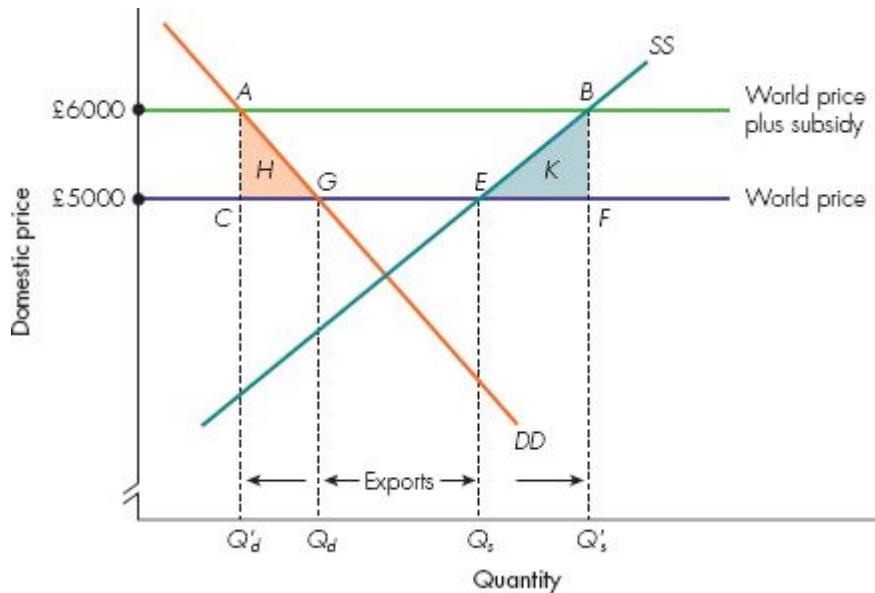


Figure 29.5 An export subsidy

Under free trade, consumers demand Q_d , production is Q_s , and exports are GE . With a subsidy on exports alone, domestic producers will restrict supply to the home market to Q'_s so that home consumers pay £6000, the same as producers can earn by exporting. Total output is Q'_s and exports AB . K shows the social cost of producing goods whose marginal cost exceeds the world price for which they are sold. H shows the social cost of restricting consumption when marginal benefits exceed the world price of the good.

To help domestic producers, the government offers a 20 per cent *export subsidy* on all exports, on which domestic producers now earn £6000. No firm sells at home for £5000 when it can sell abroad for £6000. The supply to the domestic market is reduced to Q'_s . Domestic consumers pay £6000. Total domestic output rises to Q'_s and exports are AB .

The subsidy increases exports, but entails a social cost. Triangle H is the social cost of reducing domestic consumption from Q_d to Q'_d . The consumer benefit of extra consumption would have exceeded the world

price; that is, the social marginal cost at which the economy can obtain the good. Triangle K is the social cost of raising output from Q_s to Q_s^9 when the marginal domestic cost exceeds the world price of imports.

Just as with a tariff, an export subsidy is usually a second-best policy. Even if a country wants to raise its output of computers, it is cheaper to use a production subsidy, incurring the cost of the triangle K , but avoiding the cost of the triangle H .

Summary

- World trade grew rapidly over the past 40 years, and is dominated by the **developed countries**. Primary commodities are 25 per cent of world trade; the rest is trade in manufactures.
- Countries trade because they can buy goods more cheaply abroad. Differences in costs reflect differences in technology and factor endowments. Scale economies also lead to international specialization.
- Countries make the goods in which they have a **comparative advantage** or can produce relatively cheaply. By exploiting international differences in opportunity costs, trade leads to a pure gain.
- When technology diffuses quickly to other countries, **relative factor endowments** are the main cause of different relative costs. Countries produce and export goods that use intensively the factors with which the country is relatively well endowed.
- Intra-industry trade occurs because of scale economies and consumer demand for diversity. The gain from this trade is cost reduction and greater diversity of products.
- If trade is to balance, and the forex market is to be in equilibrium, each country must have a comparative advantage in at least one good. The level of the equilibrium exchange rate offsets international differences in **absolute advantage**.

- Although international trade can benefit the world as a whole, trade usually hurts some groups of people, unless the gainers compensate the losers.
- By raising the domestic price, a **tariff** reduces consumption but raises domestic output. Hence imports fall.
- A tariff leads to two distortions that are social costs: overproduction by domestic firms whose marginal cost exceeds the world price, and under-consumption by consumers whose marginal benefit exceeds the world price.
- When a country affects the price of its imports, the world price is less than the social marginal cost of importing. This is the case for the **optimal tariff**. Otherwise, arguments for tariffs are usually second-best solutions. A production subsidy or consumption tax achieves the aim at lower social cost.
- Export subsidies raise domestic prices, reducing consumption but raising output and exports. They involve waste. Goods are exported for less than society's marginal production costs and for less than the marginal benefit to domestic consumers.
- Tariffs and other non-tariff barriers fell a lot in the last 40 years.
- Trade protection is usually costly to society. Yet governments often adopt it as an easy option politically.

Review questions



EASY

1 (a) Why does the composition of African and Asian trade differ in the table below? (b) Which pattern do you expect in the UK? Why?

Merchandise trade patterns, 2008, (% of region's exports)

	Agriculture	Fuels, minerals	Manufactures
World	8.5	22.5	66.5
North America	10.4	17	68

Europe	9.3	11.9	76.8
CIS	6.8	66.8	24.9
Africa	6.8	70.6	17.9
Middle East	2.4	74.1	21.6
Asia	6	12.4	79.2

Source: www.wto.org.

- 2 Which of the following statements is correct? (a) Now that services account for over 80 per cent of GDP in most developed countries, trade in goods cannot be the major part of the world trade. (b) International trade in services is not possible. (c) The only reason that trade in goods remains so important is that countries import goods, add a bit and then re-export them.
- 3 A country with uniformly low productivity should prevent foreign competition.— Discuss.
- 4 Common fallacies Why are these statements wrong? (a) British producers are becoming uncompetitive in everything. (b) Buy British and help Britain.
- 5 Wine, cars, steel sheeting: which have a high index of intra-industry trade? Why?

MEDIUM

- 6 Refer to the table below, which shows exports of the countries as a percentage of their GDP. Is it true that large countries gain proportionately less from world trade than small countries? Why?

	1967	2011
Belgium	36	74
Netherlands	43	73
UK	18	27
France	14	30
Italy	17	25
US	5	14
Japan	10	10

Source: *OECD, Economic Outlook*.

- 7 Usually, participating in the world economy leaves a country better-off, even though there may be winners and losers within the country. (a) Will workers with skills useful to the export industries be better or worse off when a country opens up to international trade? (b) What about workers in industries whose output is now displaced by imports? (c) Could technical progress in the export industry then ever leave a country worse off? Why, or why not?
- 8 To preserve its heritage, a country bans exports of works of art. (a) Is this better than an export tax? (b) Who gains and loses from the ban? (c) Does it encourage young domestic artists?
- 9 A small country cannot influence the price of any of the goods that it trades. What is the optimum tariff level on its imports?
- 10 Discuss the infant industry argument for a tariff on imports.
- 11 How can an export subsidy, which promotes international trade, be protectionist? Illustrate your answer with a diagram.

HARD

- 12 A perfectly competitive industry faces domestic demand $q_d = 100 - p$ and has the industry supply curve $q_s = 40 + p$. (a) If the world price is £50, what is the value of net exports? (b) If the world price is £20, what is the value of net exports? (c) In the absence of trade, what is the equilibrium domestic price? (d) At a world price of £20, suppose the government levies a tariff of £5 per unit. Calculate the value of tariff revenue and the total value of the two deadweight loss triangles.
- 13 Essay question Over the last 60 years, international trade has grown much more quickly than world output. How can this occur? Can it go on indefinitely?
- 14 Give two examples of non-tariff barriers. Draw a diagram to illustrate their effect on domestic producers and consumers.

- 1 This law was formulated by the great English economist David Ricardo (1772–1823), a successful stockbroker before retiring at the age of 40 to become a Member of Parliament and an economist. Ricardo's arguments have a modern ring to them because he used models, clearly stating their assumptions and implications.
- 2 ‘If both US goods are cheaper than UK goods when valued in pounds, they must also be cheaper when valued in dollars. We simply multiply all prices in pounds by the same exchange rate to get the corresponding dollar prices.’
- 3 ‘Strictly speaking, this explains differences in relative prices before countries start trading. Export demand may bid up the relative price of the good until relative prices are equalized across countries. This explains why trade is not infinite. Nevertheless, beginning from no

trade, comparative advantage explains which goods the country then exports and which it imports.

APPENDIX

Answers to activity and maths questions

Activity 1.1

1. In the case of the prisoner of war camp there is no production activity. Supplies were given by the Red Cross and the German Army and were the same for all prisoners. Therefore the market mechanism as outlined in the activity box does not help answering the questions ‘how to produce’ and ‘what to produce’. Those questions were resolved outside the market. The main determinant behind the creation of the market mechanism was the difference between needs and supplies. While supplies were limited, needs were not. The market could answer the question ‘for whom to produce’. Among all prisoners desiring the goods supplied, who actually receives them? The market mechanism created by pricing scarce resources ensures that a particular good is received by those who value it most. For example, if a fillet of salmon becomes a very expensive item, then only those who are willing to pay a high price for it will buy it.
2. Cigarettes play the same role as money used in exchanges. If cigarettes become scarce then fewer exchanges are to be expected. Think about yourself going into a shop with £10 to spend instead of £50. It is likely you are going to buy fewer goods. Therefore, we should expect that when cigarettes become scarce, demand for goods decreases and so do their prices. On the other hand, if cigarettes become abundant we should expect the reverse to happen and prices should increase.

Activity 2.1

1. 3.5.
2. 2.33.
3. Samantha.
4. Possible explanations include: Samantha is cleverer, concentrates more effectively and has a better memory; David’s mind is always on politics no matter how hard he studies.

Activity 3.1

- As European and Chinese labour markets are unified, the huge addition of Chinese labour makes the labour supply schedule shift to the right, reducing equilibrium wages.
- With the Chinese economy growing rapidly, the demand for raw materials such as coal is enhanced worldwide. The demand curve shifts up.
- The demand curve for Bentleys shifts down.

Maths 4.1

(a)	B	D	E
(1) Initial P and Q	$P = 2$	$P = 4$	$P = 5$
	$Q = 8$	$Q = 4$	$Q = 2$
(2) New P and Q	$P = 3$	$P = 3$	$P = 6$
	$Q = 6$	$Q = 6$	$Q = 1$
(3) % change in P	$100 * (3 - 2)/2 = 50$	$100 * (3 - 4)/4 = -25$	$100 * (6 - 5)/5 = 20$
(4) % change in Q thus induced	$100 * (6 - 8)/8 = -25$	$100 * (6 - 4)/4 = 50$	$100 * (1 - 2)/2 = -50$
(5) PED = (4)/(3)	-0.5	-2	-2.5

(b)	B	D	E
(1) Initial P and Q	$P = 2$	$P = 4$	$P = 5$
	$Q = 8$	$Q = 4$	$Q = 2$
(2) New P and Q	$P = 3$	$P = 3$	$P = 6$
	$Q = 6$	$Q = 6$	$Q = 1$
(3) % change in P	$100 * (3 - 2)/((3 + 2)/2) = 40$	$100 * (3 - 4)/((3 + 4)/2) = -28.6$	$100 * (6 - 5)/((6 + 5)/2) = 18.2$
(4) % change in Q thus induced	$100 * (6 - 8)/((6 + 8)/2) = -28.6$	$100 * (6 - 4)/((6 + 4)/2) = 40$	$100 * (1 - 2)/((1 + 2)/2) = -66.7$
(5) PED = (4)/(3)	-0.7	-1.4	-3.7

Activity 7.1

- In period 2, the cost of the machine is a sunk cost and should not enter calculation of marginal cost. Since the machine ties the two periods together, the smart way for the firm to think in period 1 is to not make a single-period decision but to make a decision over the two-period horizon, foreseeing how it will itself behave once period 2 arrives and it then has a low *MC* schedule because the machine is by then a sunk cost. Forecasting its own period-2 behaviour, it can decide in period 1 what the marginal benefit of the machine is over the two periods and choose output and investment accordingly in period 1. Tough question!

2. Sunk costs are sunk. From now on, if you think you are going to lose, ignore what you have bet and quit!

Activity 12.1

(a) Outcome	Asset price			Portfolio value		
	(a) FTSE index	(b) Low beta asset	(c) High beta asset	A 1/2 of (a) + 1/2 of (b)	B 1/3 of (a) + 2/3 of (b)	C 1/2% of (a) + 1/2 of (c)
Boom	120	90	150	105	100	135
Normal	100	100	100	100	100	100
Slump	80	110	50	95	100	65

Activity 13.1

1. Use a zero discount rate on future utility.
2. A unit of additional future consumption yields less benefit than a unit of current consumption; using a positive discount rate for future consumption reflects this. Conversely, if the burden of global warming reduces the utility of future generations below that of current generations, the consumption of future generations should then carry more weight than current generations (a negative discount rate!).
3. At least as high a rate of return on environmental investments as investment in physical or human capital.

Activity 14.1

1. Party leaders prefer the prospect of power to the adoption of a political position that they happen to think correct (and party members have elected leaders with these attributes).
2. Smart politicians will each locate in the centre.
3. Now locating too centrally risks losing some extremists, who will abstain from voting. How close a party should move to the centre depends on how many votes it loses by occupying an extreme position versus how many it gains by being closer to the centre. Political equilibrium should now have some clear blue water between the two parties, with neither contesting the exact centre ground. And of course all this presumes voter opinions are equally spread. If there is a big cluster of voters a third of the way from left to right, this is where the median voter will be. Parties will be trying to position themselves near here, even though it is not halfway from left to right.

Activity 15.1

1. It does not - that is why it is *gross* domestic product.
2. Net domestic product or net national income would include a deduction for capital depreciation.
3. Fairly rough and ready - assuming, for example, a lifetime of five years for a TV, ten years for a car, 25 years for a factory and writing the initial value off steadily over the period.
4. You would need to estimate the value of the stock of environmental capital - green fields, fresh air, temperate climate and so on - and then decide each year whether reduction caused by humans (pollution and so on) was more or less than investment made by humans (land improvement, lower emissions and so on).
5. In practice, the best way to estimate the capital value would probably be to estimate the annual consumer benefit (for example, of green fields) and then work out the present value using a suitable discount rate.

Activity 16.1

1. (a) and (c) are shifts in *AD*, but (b) is a movement along a given *AD* schedule.

Activity 17.1

1. Because fluctuations in income alter tax revenue.
2. An independent report could be useful, although regular IMF reports on individual countries perform some of the same function.
3. Bond markets would experience price falls and long-term interest rates would rise.
4. Government investment today adds to output and potential tax revenue tomorrow, and hence is close to self-financing from a long-run government perspective.
5. (i) Raise taxes and cut spending; (ii) reduces aggregate demand unless has sufficiently beneficial effect on confidence and autonomous demand to offset these effects; (iii) to the extent output falls, tax revenue will be lower than it would have been if output had not fallen; (iv) even so, the deficit will be reduced relative to what it would have been; (v) autonomous investment could go either way - depends on whether the confidence effect outweighs the short-term pessimism in relation to the immediate course of output and growth.

Activity 18.1

1. Because cash is liquid and can be used to make transactions; it is also riskless and hence may be valuable as an asset.
2. The key difference between debt and equity is that equity never has to be repaid and firms can vary the level of dividends that they pay each year - in a bad year the firm can make zero dividend payments. If all its financing was debt, it might

- (often) be unable to meet the required interest payments and have to declare bankruptcy.
3. In exchange for absorbing this degree of risk, investors in equities demand a rate of return that on average exceeds the return on debt by several percentage points a year, though because of capital gains and losses of volatile share prices, this is only true on average - particular years can be much better or much worse.
 4. Liquidity crisis. If it was a solvency crisis, the Treasury would have to be involved since taxpayers' money would then be at risk.

Activity 19.1

1. Interest rate changes will not take effect until people next recalculate their optimal behaviour - for example, it may affect new car purchases but only when people have decided that their current car is no longer appropriate.
2. Yes. As with monetary policy, it is important to distinguish changes that have immediate effect (mortgage payments, VAT changes) on disposable income, which must affect behaviour somehow, and those which operate through behavioural incentives once decision makers have recognized the effect of the new incentives.
3. People would wait to see if interest rate changes were temporary or persistent before changing big decisions such as car or house purchase.

Activity 20.1

1. 1998, 2000 and 2007.
2. Expected future inflation.
3. Lewis Hamilton does not mind turning left and then turning right shortly afterwards; however, his performance is very transparent and his skill easily assessed. Central banks need people to trust their competence, and may worry a bit more about the impact rapid reversals have on their perceived credibility. If so, they will deliberately err on the side of changing interest rates by small amounts so that the next change is likely to be in the same direction, appearing to add to their credibility. The better established their credibility, the more they might be prepared to change interest rates by a larger amount or to reverse a previous decision more quickly.
4. With a zero nominal interest rate and negative inflation, the real interest rate is positive, reducing aggregate demand, causing yet more negative inflation, yet higher real interest rates and yet lower aggregate demand.
5. A death spiral to be avoided at all costs.

Activity 21.1

1. Statistical extrapolation of past trends and cycles requires only the macroeconomic data on output itself, is quickly implemented and easily

conducted for many countries. An economist would think it a crude approximation. If there was a serious shock to the level of potential output, past extrapolation would stop working, but a statistician could always start a new extrapolation recognizing that other things were no longer equal.

2. A direct economic approach might try to use the level and rate of change of inflation to make inferences about the level of excess demand or supply, or might try to build up a more detailed model of aggregate supply based on inputs of capital, labour, technology and so on.
3. Actual output was at, or already above, potential output.

Activity 22.1

1. An Asian policy boost from monetary and fiscal policy.
2. Little.
3. The sooner Asian countries run out of spare capacity, the sooner they will have to tighten monetary policy.
4. If Japan is still concerned about deflation, it will try to increase inflation by increasing the rate of money creation, through a programme of quantitative easing and low interest rates. This resembles Europe and the US, not China and India which have continued to grow quickly and require inflation restraint.
5. They will face an adverse supply shock while demand is still low; the former will start to cause inflation and interest rates may have to rise even before output recovery is well established.

Activity 23.1

1. If you are a mathematician, try answering Review Question 11 at the end of the chapter. Otherwise, try Question 9, which will provide a diagrammatic answer. If you struggle with this, the solutions will be available to your lecturer on the OLC.
2. It does not make any difference who pays the tax. Either way, it drives the same wedge between the cost of labour to the firm and the take-home pay of the worker.
3. If there is Keynesian unemployment in the labour market, the balanced budget multiplier implies that the demand effect of extra spending on nurses outweighs the demand effect of higher taxation. If there is only equilibrium unemployment, the extra tax drives a wedge and creates extra equilibrium unemployment. Since the labour supply and job acceptances schedules are closer together at higher levels of wages and demand, adding to overall demand would in itself tend to reduce equilibrium unemployment by a small amount. On balance, the former effect might be larger, but the net result might not be large.

Activity 24.1

1. Yes.

2. Switzerland, Italy Greece, China. Switzerland, being already richest, may have least scope for future productivity growth.
3. What matters for the Balassa-Samuelson (BS) effect is traded versus non-traded. Technical progress in services makes more of these tradeable: if that was the only effect, it might not affect BS much. However, if it generally raises wages and productivity in non-tradeds, it erodes the distinction on which BS is based. As yet, we are not near that point in practice.

Activity 25.1

1. Try answering Review Question 11 at the end of the chapter. Inflation convergence took place once different countries adopted inflation targeting with similar targets.
2. Don't forget that this could also be a judgement on the weakness of UK trading partners. Optimism regarding the strength of Londons financial services could also have been significant.
3. On interest rates, not much difference. On quantitative easing, the Bank of England is more aggressive than the European Central Bank, so the UK would have suffered more. Sterling has fallen substantially and the greater competitiveness has helped mitigate the UK recession relative to what it might have been. Without this, fears about future UK growth would have been greater, and concerns about tax revenue exacerbated.

Activity 26.1

1. (i) less, (ii) less, (iii) important, (iv) less, (v) less, (vi) less.
2. Domestically if all wages and prices fall together, no real incomes change. In a closed economy the only remaining effect is that nominal debts increase in real value unless their nominal value can be adjusted too. In an open economy, cutting domestic wages and prices makes foreign goods and services more expensive - that is precisely the competitiveness gain achieved either by devaluation or by domestic price and wage cuts - in effect it is a fall in the international value of the domestic wage. Depreciation co-ordinates the international value of wage and price falls - trying to do so through domestic recession is messy and takes much longer. In itself, this is an advantage of the ability to depreciate, but may be offset by other benefits of fixed exchange rates or monetary union if other conditions are right.
3. Euro depreciation would increase export demand from outside the Eurozone, benefiting ali members to some extent. This would help peripheral Mediterranean countries but also raise the prospect of inflation within the Eurozone (higher demand and higher import prices). If the ECB reacted by raising interest rates, benefits to countries such as Greece might be lost. Greece, Spain and Portugal need especially to become more competitive vis-à-vis other EU countries. Only domestic wage and price reduction can accomplish this.

Activity 27.1

1. Wages and employment are inversely related as the economy moves along a given labour demand curve.
2. Labour demand curves shift down, which tends to mean lower wages and lower employment. If the shift is large enough, this dominates the effect in (a).
3. A supply shock leaves the demand curve unaltered and restores the answer in (a).
4. It is not worth firing to rehire in the near future. If the recession is temporary, it would be better to hang on to labour.
5. They are likely to shed labour if they become pessimistic about the length of recession.

Activity 28.1

1. Probably from two-way trade since the countries are rather similar to one another in economic structure - hence the benefits probably result more from greater competition and scale economies than from more comparative advantage. There are exceptions - for example, the agricultural Mediterranean regions probably traded different goods with Northern Europe.
2. Imperfect competition often reflects the presence of scale economies. Therefore if trade enhances market size and competition, it is probably the two-way-trade channel rather than comparative advantage that is generating gains from trade.
3. Yes, in the same way as removal of tariffs allows greater exploitation of comparative advantage. Instead of adding a tariff we are removing a barrier. The only difference is that tariffs create a revenue rectangle for government, whereas non-tariff barriers create a profit rectangle for domestic producers.

Activity 29.1

1. Probably from two-way trade since the countries are rather similar to one another in economic structure - hence the benefits probably result more from greater competition and scale economies than from more comparative advantage. There are exceptions - for example, the agricultural Mediterranean regions probably traded different goods with northern Europe.
2. Imperfect competition often reflects the presence of scale economies. Therefore if trade enhances market size and competition, it is probably the two-way-trade channel rather than comparative advantage that is generating gains from trade.
3. Yes, in the same way as removal of tariffs allows greater exploitation of comparative advantage. Instead of adding a tariff, we are removing a barrier. The only difference is that tariffs create a revenue rectangle for government, whereas non-tariff barriers create a profit rectangle for domestic .

Glossary

A

Absolute advantage an individual, company or country is the lowest-cost producer of a good.

Accelerationist hypothesis states that, by accelerating inflation faster than workers' expectations can adjust, the government can depress real wages and achieve abnormally low unemployment for a sustained period.

Accelerator model of investment assumes that firms guess future output and profits by extrapolating past output growth. Constant output growth leads to a constant level of investment. It takes accelerating output growth to raise desired investment.

Adjustable peg a fixed exchange rate, the value of which may occasionally be changed.

Aggregate demand the amount firms and households plan to spend at each level of income.

Aggregate price level measures the average price of goods and services.

Aggregate supply schedule shows the output that firms wish to supply at each inflation rate.

Appreciation (of the exchange rate) a rise in the international value of a currency.

Asset motive this motive for holding money reflects dislike of risk. People sacrifice a high average rate of return to obtain a portfolio with a lower but safer rate of return.

Asset prices jump to the level that now properly reflects new information when this becomes available.

Automatic stabilizers reduce the multiplier and thus output response to demand shocks.

Average tax rate the fraction of total income paid in tax.

B

Balance of payments records transactions between residents of one country and the rest of the world; it is the sum of current account, capital and financial account items.

Balanced budget multiplier a rise in government spending plus an equal rise in taxes leads to higher output.

Bank deposit multiplier the ratio of broad money to bank reserves.

Bank reserves the money that the bank has available to meet possible withdrawals by depositors.

Barter economy an economy in which there is no medium of exchange; goods are swapped for other goods.

Behavioural law a sensible theoretical relationship not rejected by evidence over a long period.

Benefits principle the belief that people getting most benefit from public spending should pay most for it.

Bertrand model each firm treats the prices of rivals as given.

Beta measures how much an asset's return moves with the return on the whole stock market.

Broad money includes all assets fulfilling the functions of money, and is principally bank deposits.

Budget the spending and revenue plans of an individual, company or government.

Budget constraint the different bundles that the consumer can afford.

Budget deficit the excess of government spending over government receipts.

Budget share the price of a good multiplied by its price multiplied by the quantity demanded, divided by total consumer spending or income.

Budget surplus the excess of government spending over government revenue.

Business cycle the short-term fluctuation of total output around its trend path.

C

Capital adequacy ratio a required minimum value of bank capital relative to its outstanding loans and investments.

Capital controls regulations preventing private sector capital flows between different currencies.

Capital deepening raises capital per worker for all workers.

Capital gain the rise in a share's price while it is held.

Capital widening extends the existing capital per worker to new extra workers.

Cash flow the net amount of money actually received during a period.

Central bank banker to the government and to the banks. It also conducts monetary policy.

Chosen bundle the point at which an indifference curve just touches the budget line.

Circular flow shows how real resources and financial payments flow between firms and households.

Classical model of macroeconomics assumes wages and prices are completely flexible.

Classical unemployment describes the unemployment created when the wage is deliberately maintained above the level at which the labour supply and labour demand schedules intersect.

Closed economy has no economic links with foreign countries.

Closed shop an agreement that all a firm's workers will be members of a trade union.

Collusion an explicit or implicit agreement to avoid competition.

Command economy a government planning office decides what will be produced, how it will be produced and for whom it will be produced. Detailed instructions are then issued to households, firms and workers.

Commercial banks financial intermediaries licensed to make loans and issue deposits, including deposits against which cheques can be written.

Commitment an arrangement, entered into voluntarily, that restricts future actions.

Company an organization legally allowed to produce and trade.

Comparative advantage an individual, company or country has a comparative advantage compared to another in the production of a good if they/it have (has) a lower opportunity cost in producing it.

Comparative static analysis changes one of the 'other things equal' and examines the effect on equilibrium price and quantity.

Complements goods which accompany a chosen good; a price increase for one good reduces the demand for these complements.

Constant returns to scale long-run average costs are constant as output rises.

Consumer price index (CPI) inflation measures the rate of growth of an index of consumer prices.

Consumption function shows aggregate consumption demand at each level of personal disposable income.

Contestable market has free entry and free exit.

Convergence hypothesis asserts that poor countries grow more quickly than average, but rich countries grow more slowly than average.

Convertible a currency is convertible if the central bank will buy or sell as much of the currency as people wish to trade at the fixed exchange rate.

Corporate control who controls the firm in different situations.

Corporate finance how firms finance their activities.

Cost what is spent on production during a period.

Cost of holding money the interest given up by holding money rather than bonds.

Cournot model each firm treats the output of the other firm as given.

CPI inflation measures the rate of growth of an index of consumer prices.

Credible promise a promise about the future that is optimal to carry out when the future arrives.

Credible threat one that, after the fact, is still optimal to carry out.

Credit channel affects the value of collateral for loans, and thus the supply of credit.

Cross-price elasticity of demand for good i with respect to changes in the price of good j is the percentage change in the quantity of good i demanded, divided by the corresponding percentage change in the price of good j .

Cross-section data record, at a point in time, the way an economic variable differs across different individuals or groups of individuals.

Crowding out in the Keynesian model a fiscal stimulus to aggregate demand crowds out some private spending. Higher output induces a rise in interest rates that dampens the expansionary effect on demand by reducing some components of private spending.

Crowding out in the classical model a rise in government spending crowds out an equal amount of private spending. Aggregate demand remains equal to potential output.

Currency board a constitutional commitment to peg the exchange rate by giving up monetary independence.

Current account of the balance of payments records international flows of goods, services and current transfers.

Current transfers transfer payments paid across borders.

D

Data pieces of evidence about economic behaviour.

Deadweight loss lost social surplus by producing the wrong output level.

Deflation negative inflation, when the price level is falling.

Deflationary gap measures the shortfall of aggregate demand over output when output is at potential output.

Demand the quantity that buyers wish to purchase at each conceivable price.

Demand curve shows the relationship between price and quantity demanded, other things equal.

Demand-deficient unemployment occurs when output is below full capacity.

Demand-determined output since markets trade the smaller of supply and demand, output is demand-determined when there is excess supply and wages and prices have yet to adjust to restore long-run equilibrium. Output then depends only on aggregate demand.

Demand elasticity demand is elastic if the price elasticity is more negative than – 1.

Demand for money a demand for real money balances.

Demand inelasticity demand is inelastic if the price elasticity lies between – 1 and 0.

Demand management uses monetary and fiscal policy to stabilize output near potential output.

Demand shocks when all shocks are demand shocks, stabilizing inflation also stabilizes output, even in a Keynesian model.

Depletable resources resources that can be used only once.

Depreciation the loss in value of a capital good during a period.

Depreciation (of the exchange rate) a fall in the international value of a currency.

Derived demand the demand for inputs reflects demand for a firm's output.

Devaluation (revaluation) reduces (increases) the par value of the pegged exchange rate.

Diminishing marginal rate of substitution tastes exhibit this when, to hold utility constant, diminishing quantities of one good must be sacrificed to get successive equal increases in the quantity of the other good.

Diminishing marginal utility each extra unit consumed, holding constant consumption of other goods, adds successively less to total utility.

Direct taxes taxes on income and wealth.

Discount rate the interest rate that the Bank charges when banks want to borrow cash.

Discouraged workers people pessimistic about finding a job who leave the labour force as a result.

Discretion free choice without restrictions imposed by prior commitments.

Discretionary fiscal policy decisions about tax rates and levels of government spending.

Discriminating monopoly one which charges different prices to different people.

Diseconomies of scale (or decreasing returns to scale) longrun average cost rises as output rises.

Disposable income is gross income minus taxes plus benefits; that is, the net income available to spend or save.

Distortion exists if society's marginal cost of producing a good does not equal society's marginal benefit from consuming that good.

Diversification pools risk across several assets whose individual returns behave differently from one another.

Dividends the regular payments of profit to shareholders.

Domestic price of foreign exchange quantity of domestic currency per unit of foreign currency.

Dominant strategy a player's best strategy *whatever* the strategies adopted by rivals.

Dumping occurs when foreign producers sell at prices below their marginal production cost, either by making losses or with the assistance of government subsidies.

E

Econometrics is the term used to refer to statistics to measure relationships in economic data.

Economic growth a rise in real GNP.

Economic rent (not to be confused with income from renting out property) is the payment a factor receives in excess of what is needed to keep it in its present occupation.

Economic sovereignty the power of national governments to make decisions independently of those made by other governments.

Economics the study of how society decides what, how and for whom to produce.

Economies of scale (or **increasing returns to scale**) long-run average cost falls as output rises.

Effective exchange rate an average of a country's exchange rate against all its trade partners, weighted by the relative size of trade with each country.

Efficiency wages high wages that raise productivity through their incentive effect.

Efficient asset market already incorporates existing information properly in asset prices.

Endogenous growth implies that the steady-state growth rate is affected by economic behaviour and economic policy.

Endogenous variables those variables which a model delivers as outputs, conditional on the values of the exogenous inputs.

Entry when new firms join an industry.

Equilibrium output is independent of inflation.

Equilibrium price the price at which the quantity supplied equals the quantity demanded.

Equilibrium unemployment (also called the natural rate of unemployment) the unemployment rate when the labour market is in equilibrium.

Excess demand exists when the quantity demanded exceeds the quantity supplied at the ruling price.

Excess supply exists when the quantity supplied exceeds the quantity demanded at the ruling price.

Exchange rate the price at which two currencies exchange.

Exchange Rate Mechanism (ERM) part of the EMS. Each country fixed a nominal exchange rate against each other ERM participant. Collectively, the group floated against the rest of the world.

Exchange rate regime describes how governments allow exchange rates to be determined.

Exit when existing firms leave an industry.

Exogenous variables those variables fed into a model as inputs.

Export subsidies government assistance to domestic firms in competing with foreign firms.

Exports domestically produced goods that are sold abroad.

External balance a zero current account balance.

Externality arises if one person's production or consumption physically affects the production or consumption of others.

Extrapolative expectations assume that the future is an extension of the recent past.

Extreme Keynesians believe markets do not clear, even in the long run.

F

Fair gamble a bet which, on average, yields zero monetary profit.

Fallacy of composition what is true for the individual may not be true for everyone together, and what is true for everyone together may not hold for the individual.

Fan chart indicates the probability of different outcomes.

Federal fiscal system has a central government setting taxes and expenditure rules that apply in its constituent states or countries.

Final goods goods purchased by the ultimate user, either households buying consumer goods or firms buying capital goods such as machinery.

Financial account of the balance of payments records international purchases and sales of financial assets.

Financial intermediary specializes in bringing lenders and borrowers together.

Financial panic a self-fulfilling prophecy in which, believing a bank will be unable to pay, people rush to get their money out. But doing so makes the bank bankrupt.

First-best allocation is fully efficient; it removes all distortions.

First-mover advantage the player moving first achieves higher payoffs than when decisions are simultaneous.

Fiscal drag the rise in real tax revenue when inflation raises nominal incomes, pushing people into higher tax brackets in a progressive income tax system.

Fiscal policy government policy on spending and taxes.

Fiscal stance shows the effect of fiscal policy on demand and output.

Fisher hypothesis says higher inflation leads to similarly higher nominal interest rates.

Fixed costs total costs do not vary with output.

Fixed exchange rate regime governments maintain the convertibility of their currency at a fixed exchange rate. A currency is convertible if the central bank will buy or sell as much of the currency as people wish to trade at the fixed exchange rate.

Fixed factor of production an input that cannot be varied.

Flexible inflation targeting commits a central bank to hit inflation targets in the medium run, but gives it some discretion about how quickly to hit its inflation target.

Flight from cash the collapse in the demand for real cash when high inflation and high nominal interest rates make it very expensive to hold cash.

Floating exchange rates the exchange rate is allowed to find its equilibrium level without central bank intervention using the forex reserves.

Flow the stream of accounts measured over a period of time.

Foreign direct investment (FDI) the purchase of foreign firms or the establishment of foreign subsidiaries.

Foreign exchange (forex) market exchanges one national currency for another.

Foreign exchange reserves foreign currency held by a domestic central bank.

Forward market deals in contracts made today for delivery of goods at a specified future date at a price agreed today.

Free markets markets in which governments do not intervene.

Free rider a person who, unable to be excluded from consuming a good, has no incentive to buy it.

Frictional unemployment the irreducible minimum unemployment in a dynamic society.

Functional income distribution the division of national income among different factors of production.

G

Game a situation in which intelligent decisions are necessarily interdependent.

GDP at basic prices measures domestic output exclusive of indirect taxes on goods and services.

GDP at market prices measures domestic output inclusive of indirect taxes on goods and services.

GDP deflator is the ratio of nominal GDP to real GDP expressed as an index.

Globalization the increasing integration of national markets that were previously much more segmented from one another.

GNP deflator the ratio of nominal GNP to real GNP expressed as an index.

Government solvency requires that the present value of the current and future tax revenue equals the present value of current and future spending plus any initial net debts.

Gradualist monetarists believe that full employment is restored within a few years, so the main effect of higher money is higher prices.

Gross debt is total liabilities owed to creditors.

Gross domestic product (GDP) the value of total output of an economy in a given period.

Gross investment the production of new capital goods and the improvement of existing capital goods.

Gross national product (GNP) measures the income of an economy; the quantity of goods and services the economy can afford to purchase.

Growth rate the percentage change per period (usually a year).

H

Headline inflation actual inflation reflected by growth in the retail price index.

Hedging the use of forward markets to shift risk on to somebody else.

Higher government spending on goods and services increases equilibrium output.

Higher net tax rate reduces both equilibrium output and the budget deficit for given government spending G.

Horizontal equity the identical treatment of identical people.

Horizontal LM schedule implies the money supply is adjusted to keep interest rates constant.

Hyperinflation a period of very high inflation.

Hysteresis when a particular long-run equilibrium depends on the path any economy follows in the short run.

I

ii schedule shows that, at higher inflation rates, the central bank will wish to have higher real interest rates.

Imperfectly competitive firm a firm that faces a downward-sloping demand curve. Its output price reflects the quantity of goods it makes and sells.

Import tariff a tax on imports.

Imports goods produced abroad but purchased for use in the domestic economy.

Impossible triad fixed exchange rates, perfect capital mobility and monetary sovereignty. All three cannot co-exist at the same time.

Income distribution tells us how total income is divided between different groups or individuals (in a country or in the world).

Income effect the income effect of a price change is the adjustment of demand to the change in real income alone.

Income elasticity of demand the percentage change in quantity demanded of a good divided by the corresponding percentage change in income.

Income expansion path how the chosen bundle of goods varies with consumer income levels, keeping constant everything else.

Income policy the direct control of wages and other income.

Index number expresses data relative to a given base value.

Indifference curve shows all the consumption bundles yielding a particular level of utility.

Indirect taxes taxes on spending and output.

Inferior good a good for which demand falls when incomes rise; it has a negative income elasticity of demand.

Inflation a rise in the price level.

Inflation accounting uses fully inflation-adjusted definitions of costs, income and profit.

Inflation illusion confusing nominal and real changes. People's welfare depends on real variables, not nominal variables.

Inflation rate the percentage increase in the average price of goods and services.

Inflation target the central bank adjusts interest rates to try to keep inflation close to the target inflation rate.

Inflation tax the effect of inflation in raising real revenue by reducing the real value of the government's nominal debt.

Inflation-adjusted budget uses real not nominal interest rates to calculate government spending on debt interest.

Inflationary gap measures the excess of aggregate demand over output when output is at potential output.

Injection money that flows to firms without being recycled through households.

Innocent entry barrier one not deliberately erected by incumbent firms.

Innovation the incorporation of new knowledge into actual production techniques.

Input (or **factor of production**) a good or service used to produce output.

Insider those with jobs, represented in wage bargaining.

Interest parity means that expected exchange rate changes offset the interest differential between domestic and foreign currency assets.

Interest rate spread the excess of a loan interest rate over a deposit interest rate.

Intermediate goods goods that are partly finished and form inputs to a subsequent production process that then uses them up.

Intermediate target a key indicator used to guide interest rate decisions.

Internal balance aggregate demand equals potential output.

International monetary system provides a medium of exchange for international transactions.

International value of the domestic currency quantity of foreign currency per unit of domestic currency.

Intra-industry trade two-way trade in goods made within the same industry.

Invention the discovery of new knowledge.

Inventories goods held in stock by the firm for future sales.

Investment the purchase of new capital goods by firms.

Investment demand firms' desired or planned additions to physical capital (factories and machines) and to inventories.

Investment demand schedule shows desired investment at each interest rate.

Invisible hand the assertion that the individual pursuit of self-interest within free markets may allocate resources efficiently from society's viewpoint.

Involuntary unemployment when workers want to work at the going wage but cannot find jobs.

IOU money a medium of exchange based on the debt of a private firm or individual.

IS schedule shows combinations of income and interest rates at which aggregate demand equals actual output.

Isoquant shows minimum combinations of inputs to make a given output. Different points on an isoquant reflect different production techniques.

L

Labour-augmenting technical progress increases the effective labour supply.

Labour force all individuals in work or looking for work.

Laffer curve shows how much tax revenue is raised at each possible tax rate.

Land the factor of production that nature supplies.

Law of comparative advantage countries specialize in producing and exporting the goods that they produce at a lower relative cost compared to other countries.

Law of diminishing marginal returns further increases in a variable input lead to steadily decreasing marginal product of that input.

Law of diminishing returns says each extra worker adds less to output than the previous extra worker added.

Leakage leakage from the circular flow is money not recycled from households to firms.

Lender of last resort the central bank: it lends to banks when financial panic threatens the financial system.

Less-developed countries (LDCs) countries with low per capita incomes.

Life-cycle hypothesis assumes people make a lifetime consumption plan (including bequests to their children) that is just affordable out of lifetime income (plus any initial wealth inherited).

Limited liability shareholders of a company cannot lose more than they have already invested in the business.

Liquidity the cheapness, speed and certainty with which asset values can be converted back into money.

Liquidity crisis an institution is temporarily unable to meet immediate requests for payment even though its underlying assets exceed its liabilities.

LM schedule shows combinations of interest rates and income yielding money market equilibrium when the central bank pursues a given target for the nominal money supply.

Log-rolling a vote for another person's preferred outcome on one issue in order to exchange for their vote your preferred outcome on another issue.

Long run the period needed for complete adjustment to a price change; its length depends on the type of adjustments consumers wish to make.

Long-run average cost the total cost (*LTC*) divided by level of output.

Long-run equilibrium when the price equates the quantity demanded to the total quantity supplied by the number of firms in the industry when each firm is on its long-run supply curve and firms can freely enter or exit the industry.

Long-run marginal cost the rise in long-run total cost if output rises permanently by one unit.

Long-run supply curve how price affects desired output; hence, the part of a firm's *LMC* curve above its *LAC* curve.

Long-run total cost the minimum cost of producing each output level when a firm can adjust all inputs.

Long-term interest rates apply to long-term loans during which the interest rate is usually fixed.

Luxury good a good with an income elasticity above unity.

M

Maastricht criteria for joining EMU said that a country must already have achieved low inflation and sound fiscal policy.

Macroeconomics the study of the economy as a system.

Managed float central banks intervene in the forex market to try to smooth out fluctuations and nudge the exchange rate in the desired direction.

Marginal cost the rise in total cost when output rises 1 unit.

Marginal firm the last firm to enter in the market, which makes zero long-run profits.

Marginal product the marginal product of a variable factor is the extra output from an extra unit of that input, holding constant all other inputs.

Marginal product of labour is the extra total output when an extra worker is added, with other input quantities unaltered.

Marginal propensity to consume the fraction of each extra pound of disposable income that households wish to consume.

Marginal propensity to import (MPZ) the fraction of each extra pound of national income that domestic residents wish to spend on extra imports.

Marginal propensity to save the fraction of each extra unit of income that households wish to save.

Marginal rate of substitution the quantity of a good a consumer must sacrifice to increase the quantity of another good by one unit without changing total utility.

Marginal revenue the rise in total revenue when output rises 1 unit.

Marginal revenue product of labour the change in total output revenue when a firm sells the extra goods that an extra unit of labour input allows it to produce.

Marginal tax rate the fraction of the last pound of income paid in tax.

Marginal utility of a good is the increase in total utility obtained by consuming one more unit of that good, for a given consumption of other goods.

Marginal value product of capital the extra value of the firm's output when another unit of capital services is used, all other inputs being held fixed.

Marginal value product of labour the extra revenue from selling the output made by an extra worker.

Market a process by which households' decisions about consumption of alternative goods, firms' decisions about what and how to produce, and workers' decisions about how much and for whom to work are all reconciled by adjustment of prices.

Market demand curve the sum of the demand curves of all individuals in that market.

Median voter the person whose opinion represents the middle position on an issue.

Medium of exchange something accepted as payment only to be subsequently reused to pay for something else.

Menu costs of inflation the physical resources needed for adjustments to keep real things constant when inflation occurs.

Merit (demerit) goods goods that society thinks everyone should have (not have) regardless of whether an individual wants them.

Microeconomics offers a detailed treatment of how individuals and firms make economic decisions.

Minimum efficient scale (MES) the lowest output at which a firm's *LAC* curve stops falling.

Mismatch occurs if the skills that firms demand differ from the skills the labour force possesses.

Mixed economy a system in which the government and private sector jointly solve economic problems. The government influences decisions through taxation, subsidies and provision of free services such as defence and the police. It also regulates the extent to which individuals may pursue their own self-interest.

Model or theory makes assumptions from which it deduces how people will behave. It is a deliberate simplification of reality.

Moderate Keynesians believe the economy will eventually return to full employment, but that this could take many years.

Monetary base (or **narrow money**) the quantity of notes and coins in private circulation plus the quantity of reserves held by commercial banks.

Monetary instrument the variable over which the central bank makes day-to-day choices.

Monetary policy a relationship between the state of the economy and the interest rate chosen by the central bank.

Monetary policy accommodates a temporary supply shock when monetary policy is altered to help stabilize output. The consequence,

however, is higher inflation.

Monetary sovereignty a country's monetary independence. It is undermined if interest rates are set to maintain the pegged exchange rate because they cannot be set independently to influence the domestic economy.

Monetary target adjusting interest rates to maintain the quantity of money demanded in line with a given target for money supply.

Monetary union a commitment to permanently fixed exchange rates, an integrated financial market and a single central bank setting the single interest rate for a union.

Money any generally accepted means of payment for delivery of goods or settlement of debt.

Money illusion exists if people confuse nominal and real variables.

Money market equilibrium a situation in which the quantity of real balances demanded and supplied is equal.

Money multiplier the ratio of broad money to the monetary base.

Money supply currency in circulation outside the banking system, plus deposits of commercial banks and building societies.

Monopolist the only seller or potential seller in the industry.

Monopolistic competition an industry with many sellers of products that are close substitutes for one another. Each firm has only a limited ability to affect its output price.

Monopoly power power exhibited by the excess of price over marginal cost.

Monopsony power with an upward-sloping factor supply curve, a firm must offer a higher factor price to attract more factors. The marginal cost of the input exceeds the factor price, since the firm bids up the price paid on all inputs already employed.

Multinationals firms operating in many countries simultaneously.

Multiplier the ratio of a change in equilibrium output to a change in autonomous spending that caused the change.

N

Nash equilibrium each player chooses the best strategy, *given* the strategies being followed by other players.

National debt the government's debts.

Natural level of output the level of output in long-run equilibrium.

Natural monopoly has falling average cost no matter how high its output rises. It undercuts all smaller competitors and fears no entrant.

Natural rate of unemployment the level of unemployment in long-run equilibrium.

Necessity a good with an income elasticity below unity.

Net debt is total liabilities minus total assets that could be sold in order to raise money to pay creditors.

Net investment gross investment minus depreciation of the existing capital stock.

Net taxes taxes minus transfers.

Net worth assets a firm owns minus liabilities it owes.

New Classical macroeconomics is based on the twin principles of rapid market clearing and rational expectations.

New Keynesians provide rigorous microeconomic foundations for macroeconomics with temporary price rigidity.

N-firm concentration ratio the market share of the largest N firms in the industry.

Nominal anchor determines the level of other nominal variables. Market forces determine real variables.

Nominal GDP measures GDP at the prices prevailing when output was produced.

Nominal GNP measures GNP at the prices prevailing when income was earned.

Nominal interest rate how many actual pounds are earned by lending £1 for a year.

Nominal money growth in the classical model faster nominal money growth is accompanied by higher inflation but leaves real output constant at potential output.

Nominal values are measured in the prices ruling at the time of measurement.

Non-tariff barriers different national regulations or practices that prevent free movement of goods, services and factors between countries.

Normal good a good for which demand increases when incomes rise; it has a positive income elasticity of demand.

Normal profits accounting profits which just cover the opportunity cost of an owner's money and time.

Normative economics offers recommendations based on personal value judgements.

O

Oligopoly an industry with few producers, each recognizing their interdependence.

Open economy an economy with important trade and financial links with other countries.

Open economy macroeconomics examines how the economy is affected by links with other countries through trade, the exchange rate and capital flows.

Open market operation occurs when the central bank alters the monetary base by buying or selling financial securities in the open market.

Opportunity cost of an activity is the value of the best alternative you must sacrifice.

Optimal currency area a group of countries that is better off with a common currency than keeping to separate national currencies.

Optimal tariff a tariff to reduce imports to the level at which social marginal cost equals social marginal benefit.

Other things equal a device for looking at the relationship between two variables, but remembering other variables also matter.

Output gap the deviation of actual output from potential output.

Outsiders those without jobs, who are unrepresented in wage bargaining.

P

Panel data record observations over multiple time periods for the same individuals or groups of individuals.

Par value the exchange rate that the government agrees to defend.

Paradox of thrift a change in the amount households wish to save at each income leads to a change in equilibrium income, but no change in equilibrium saving, which must still equal planned investment.

Pareto efficiency for given tastes, resources and technology, an allocation is efficient if there is no other feasible allocation that makes some people better off and nobody worse off.

Participation rate the fraction of the population of working age who join the labour force.

Partnership a business jointly owned by two or more people, sharing the profits and being jointly responsible for any losses.

Per capita real GNP real GNP divided by the total population. It is real GNP per head.

Percentage change the absolute change divided by the original number, then multiplied by 100.

Perfect capital mobility means that a vast quantity of funds flow from one currency to another if the expected return on assets differs across currencies.

Perfect competition a market in which both buyers and sellers believe that their own actions have no effect on the market price.

Permanent income hypothesis a belief that consumption reflects long-run or permanent income.

Permanent supply shock changes potential output.

Personal disposable income the income households receive from firms, plus transfer payments received from the government, minus direct taxes paid to the government. It is the net income households can spend or save.

Personal income distribution the division of national income across individuals, regardless of the factor services from which these individuals earn their income.

Phillips curve shows that a higher inflation rate is accompanied by a lower unemployment rate. It suggests we can trade off more inflation for less unemployment or vice versa.

Physical capital machinery, equipment and buildings used in production.

Policy co-ordination the decision to set policies jointly when two interdependent areas have big cross-border spillovers.

Political business cycle arises if politicians manipulate the economy for electoral advantage.

Political economy the study of how governments make decisions.

Positive economics studies objective or scientific explanations of how the economy works.

Potential output the economy's output when inputs are fully employed.

Precautionary motive holding money to meet contingencies that we cannot yet foresee.

Present value the present value of a future £1 is the sum that, if lent today, would cumulate to £1 by that date.

Price controls government rules or laws setting price floors or ceilings that forbid the adjustment of prices to clear markets.

Price elasticity of demand (PED) the percentage change in the quantity demanded divided by the corresponding percentage change in its price. $PED = (\% \text{ change in quantity}) / (\% \text{ change in price})$.

Price of an asset the sum for which an asset can be purchased outright. The owner of a capital asset gets the future stream of capital services from that asset.

Principal-agent problem difficulties of a principal or owner in monitoring an agent to whom decisions have been delegated.

Principle of targeting the most efficient way to attain a given objective is to use a policy influencing that activity directly.

Private good a good that, if consumed by one person, cannot be consumed by others.

Production efficiency means more output of one good can be obtained only by sacrificing output of other goods.

Production function the maximum output obtainable from specified quantities of inputs, given existing technical knowledge.

Production possibility frontier (PPF) shows, for each output of one good, the maximum amount of the other good that can be produced.

Profit revenue minus cost.

Property rights the power of residual control, including the right to be compensated for externalities.

Public good a good that, if consumed by one person, must be consumed by others in exactly the same quantity.

Purchasing power of money an index of the quantity of goods that can be bought for £1.

Purchasing power parity (PPP) the path of the nominal exchange rate that would maintain a constant real exchange rate. Nominal exchange rate changes offset inflation differentials between countries.

Pure inflation prices of goods and inputs rise at the same rate.

Q

Quantitative easing the creation of substantial quantities of bank reserves in order to offset a fall in the bank deposit multiplier and prevent large falls in bank lending and broad money.

Quantity theory of money changes in nominal money lead to equivalent changes in the price level (and money wages), but have no effect on output and employment.

Quotas restrictions on the maximum quantity of imports.

R

Rational expectations guess the future correctly on average.

Reaction function how optimal actions by one player vary with the assumed actions of the other player.

Real business cycle theories explain cycles as fluctuations in potential output itself.

Real exchange rate the relative price of goods from different countries when measured in a common currency.

Real GDP, or GDP at constant prices, adjusts for inflation, measuring GDP in different years at the prices prevailing at a particular date, known as the *base year*.

Real GNP adjusts for inflation by measuring GNP in different years at the prices prevailing at some particular date known as the base year.

Real interest rate the return on a loan, adjusted for inflation, which shows as the extra quantity of goods earned by postponing consumption.

Real money supply the nominal money supply M divided by the price level P .

Real values adjust nominal values for changes in the price level.

Regulatory capture the regulator gradually comes to identify with the interests of the firm it regulates, becoming its champion not its watchdog.

Renewable resources resources that can be used again if not over-exploited.

Rental rate (for capital) the cost of using capital services.

Replacement rate the level of benefits relative to wages in work.

Repo the sale of an asset with a simultaneous agreement to repurchase later.

Required rental on capital an amount which just covers the opportunity cost of owning the asset.

Required reserve ratio a minimum ratio of cash reserves to deposits that banks are required to hold.

Reservation wage is the lower wage a worker is willing to accept to work in a given occupation.

Reserve ratio the ratio of reserves to deposits.

Resource allocation a complete description of who does what and who gets what.

Retained earnings the part of after-tax profits ploughed back into a business.

Revenue what the firm earns from selling goods or services in a given period.

Reverse repo a purchase with a simultaneous agreement to resell later.

Ricardian equivalence says that it does not matter when a government finances a given spending programme. Tax cuts today do not affect private spending if, in present value terms, future taxes rise to match.

Risk-averse a person who will refuse a fair gamble.

Risk-lover a person who bets even when the odds are unfavourable.

Risk-neutral a person who is interested only in whether the odds yield a profit on average.

Risk pooling aggregates independent risks to make the aggregate more certain.

Risk sharing works by reducing the stake.

Rule a commitment describing how behaviour changes when circumstances change.

S

Saving the part of income not spent buying goods and services.

Saving function shows desired saving at each income level.

Scarce resource a resource is scarce if the demand of that resource at a zero price would exceed the available supply.

Scatter diagram plots pairs of values simultaneously observed for two different variables.

Second-best the most efficient outcome that can be achieved conditional on being unable to remove some distortions.

Seigniorage real revenue acquired by the government through its ability to print money.

Self-fulfilling prophecy an expectation that creates the incentive to make it come true.

Shoe-leather costs of inflation the extra time and effort in transacting when we economize on holding real money.

Short run the period after prices change but before quantity adjustment can occur.

Short-run average fixed cost(SAFC) short-run fixed cost (SFC) divided by output.

Short-run average total cost (SATC) short-run total cost (STC) divided by output.

Short-run average variable cost (SAFC) short-run variable cost (SVC) divided by output.

Short-run equilibrium when the price equates the quantity demanded to the total quantity supplied by the given number of firms in the industry when each firm is on its short-run supply curve.

Short-run equilibrium output the output at which aggregate demand or planned spending equals the output actually produced.

Short-run marginal cost the extra cost of making an extra unit of output in the short run while some inputs remain fixed.

Short-run output decision the firm supplies the output at which $MR = SMC$, if the price covers average variable cost $SAVC_1$ at that output. If not, the firm supplies zero.

Short-run Phillips curve higher unemployment is associated with lower inflation. The height of the short-run Phillips curve reflects expected inflation.

Short-run supply curve (SAS) shows how desired output varies with inflation, for a given inherited growth of nominal wages.

Short-term interest rates apply to loans of very short maturity.

Shutdown price the price below which the firm cuts its losses by making no output.

Sight deposits money that can be withdrawn ‘on sight’ without prior notice.

Social cost of monopoly the failure to maximize social surplus.

Sole trader a business owned by a single individual.

Solow residual the part of output growth not explained by the growth of measured inputs.

Solvency crisis an institution’s assets have become less than its liabilities. The institution is then bankrupt without a rapid new injection of assets from government or shareholders.

Speculation the purchase of an asset for subsequent resale, in the belief that the total return - interest plus capital gain - exceeds the total return on other assets.

Speculative attack a large capital outflow. If successful, it causes a devaluation. Attacks are sometimes resisted, by raising interest rates and tightening fiscal policy. This works only if it can credibly be sustained.

Speculator a person who temporarily holds an asset in the hope of making a capital gain.

Spot market deals in contracts for immediate delivery and payment.

Stabilization policy government action to keep output close to potential output.

Stackelberg model a firm with a first-mover advantage can deduce how its actions induce rivals subsequently to behave.

Stagflation high inflation and high unemployment, caused by an adverse supply shock.

Steady-state path output, capital and labour grow at the same rate. Hence output per worker and capital per worker are constant.

Sterilization an open market operation between domestic money and domestic bonds, to offset the change in domestic money supply that a

balance of payments surplus or deficit otherwise induces.

Sterilized intervention does not work under perfect capital mobility

because offsetting capital flows are immediately induced.

Stock a quantity at a given point in time.

Store of value any asset whose value largely lasts into the next period.

Strategic entry deterrence behaviour by incumbent firms to make entry of other firms less likely.

Strategic move a move that influences the other person's choice, in a manner favourable to oneself, by affecting the other person's expectations of how you will behave.

Strategy a game plan describing how a player acts, or moves, in each possible situation.

Structural budget shows what the budget will be if output is at potential output.

Structural unemployment arises from the mismatch of skills and job opportunities as the pattern of demand and supply changes.

Substitutes alternative goods sought by consumers as a result of a price increase in their original choice of good; a price rise in the chosen good raises the demand for these substitutes.

Substitution effect the substitution effect of a price change is the adjustment of demand to the relative price change alone.

Supernormal profit pure economic profit after measuring all economic costs properly.

Supply the quantity of a good that sellers wish to sell at each possible price.

Supply curve shows the relationship between price and quantity supplied, other things equal.

Supply-determined output output determined by excess demand, as under rationing.

Supply-side economics (1) analyses how taxes and other incentives affect national output when the economy is at full capacity; (2) the use of microeconomic incentives to alter the level of full employment, the level of potential output and equilibrium unemployment.

T

Tangency equilibrium in the long run, each firm's demand curve just touches its *AC* curve at the output level at which *MC* equals *MR*. Each firm maximizes profits but just breaks even. There is no more entry or exit.

Tangible wealth capital and land.

Tariff a tax on imports.

Tax incidence describes who eventually bears the burden of that tax.

Tax wedges are created when the marginal tax rate is the fraction of each extra pound that the government takes in tax. Tax wedges occur between the price the purchaser pays and the price the seller receives.

Technical efficiency exists if there is no other way to make a given output using less of one input and no more of the other inputs.

Technical progress a new technique allowing a given output to be made with fewer inputs than before.

Temporary supply shock shifts the short-run aggregate supply schedule, but leaves potential output unaltered.

Time deposits these deposits, paying higher interest rates, require the depositor to give notice before withdrawing money.

Time series a sequence of measurements of the same variable at different points in time.

Tobin tax a small tax on capital flow transactions.

Token money a means of payment whose value or purchasing power as money greatly exceeds its cost of production or value in uses other than as money.

Trade balance the value of net exports.

Trade deficit net exports are negative.

Trade policy policy that affects international trade through taxes or subsidies, or by direct restrictions on imports and exports.

Trade surplus net exports are positive.

Trade union power is measured by the ability of unions to co-ordinate lower job acceptances, thereby increasing wages but reducing employment.

Transactions motive the transactions motive for holding money reflects the fact that payments and receipts are not synchronized.

Transfer earnings the minimum payments needed to induce the input to work in that job.

Transfer in kind the gift of a good or service.

Transfer payment a payment, usually by the government, for which no corresponding service is provided by the recipient.

Transmission mechanism of monetary policy is the channel through which it affects output and employment.

Trend path of output the smooth path of long-run output once its short-term fluctuations are averaged out.

Two-part tariff a fixed charge for access to a service and then a price per unit that reflects the marginal cost of production.

U

Underlying inflation growth in the retail price index, after omitting the effect of mortgage interest rates on the cost of living.

Unemployment rate the fraction of the labour force without a job but registered as looking for work.

Unit elasticity demand elasticity is – 1.

Unit of account the unit in which prices are quoted and accounts kept.

Unsterilized intervention uses forex reserves to offset balance of payments surpluses or deficits. Since foreign reserves are exchanged for domestic cash, this alters the cash in circulation and the domestic money supply.

V

Value added the increase in the value of goods as a result of the production process.

Variable costs total costs change with output.

Variable factor an input that can be varied, even in the short run.

Vertical equity the different treatment of different people in order to reduce the consequences of these innate differences.

Voluntary unemployment when, at the given level of wages, a worker wishes to be in the labour force but does not yet wish to accept a job.

W

Wealth effect the shift in the consumption function when household wealth changes.

Welfare economics deals with normative issues. It does not describe how the economy works but assesses how well it works.

Z

Zero-growth proposal because higher measured GNP imposes environmental costs, it is best to aim for zero growth of measured GNP.

INDEX

A

- absolute advantage 12–13
 - see also* comparative advantage
- international trade 658
- opportunity cost 12–13
- acceleration hypothesis, expectations and credibility 505–6
- accelerator model of investment, business cycles 611–13
- accounting costs, firms' accounts 120–1
- accounting profits vs economic profits 121
- accumulation, economic growth 638–41
- actual output/potential output, macroeconomics 367–8
- addresses, best 262
- adjustable peg, exchange rate regimes 587
- adjustment in the market, capital services 259–61
- adjustment process
 - aggregate demand schedule 484–5
 - aggregate supply schedule 484–5
- adverse selection
 - asymmetric information 276–7
 - risk 276–7
- aggregate demand 365–86
 - aggregate demand schedule 373
 - components 368–71
 - confidence 383–4
 - consumption demand 368–370
 - consumption function 369–70
 - endogenous variables 371
 - equilibrium 373–7
 - equilibrium output 372–6
 - exogenous variables 371
 - fall in aggregate demand 377–80
 - fiscal stability 401–3
 - fixed exchange rates 573–4
 - government 388–93
 - ii schedule 476
 - inflation 474–5
 - investment demand 371
 - investment spending 371

IS schedule 476
multiplier 380–2
output 365–86
paradox of thrift 381–3
personal disposable income 368–9
planned saving/planned investment 376–7
saving rates 382–3
short-run equilibrium output 372
aggregate demand schedule 373, 475–6
adjustment process 484–5
aggregate price level 18
aggregate supply 477
credit constraints 617
Phillips curve 502–3
aggregate supply schedule 477
adjustment process 484–5
agricultural protection, international trade 656
anticipated inflation 508–10
appreciation/depreciation, foreign exchange market 544
apps, mobile, demand 48
Asia, controlling inflation 513–14
Asian tigers, economic activity 359–60
asset motive, holding money 424
asset price 250
asset returns
see also investment
asset returns correlation, diversification 281–2
beta 282–4
efficient asset markets 284–7
risk 277–9
uncertainty 277–9
asset valuation 252
assets, capital *see* capital assets
asymmetric information
adverse selection 276–7
insurance market 276–7
market failure 303
moral hazard 275–6
risk 275–7
attitudes, individual, risk 269–71
austerity
fiscal policy 391–3
Japan 391–3
automatic stabilizers, discretionary fiscal policy 398–9
average tax rate 321
averages, index numbers 24

B

balance of payments 547–8
 crises 590–1
 exchange rate regimes 548
 fixed exchange rates 548
 floating exchange rates 548
 money supply 564–5

balance of payments adjustment, gold standard 585–6

balance sheet, firms' accounts 119–20

balanced budget multiplier 390

bank deposit multiplier 419–22

Bank of England 434–5
 see also central banks
 controlling inflation 513
 independence 513
 inflation 479–513
 open market operations 436

bank reserves 416

banking 413–33
 see also central banks; financial markets; money
 bank deposit multiplier 419–22
 bank reserves 416
 banking crises, preventing 438–9
 Basel III Accord 439
 broad money 421–423
 business of banking 416
 capital adequacy ratios 437–9
 cash reserves 416
 central banks 419
 clearing system 415–16
 collapse of bank lending 421–2
 commercial banks 416
 competition 441
 cost of holding money 423
 creating money 416–17
 demand for money 423–6
 discount rate 436
 financial crisis (2007–81) 427–31
 financial intermediaries 416
 government bailout 438–9
 holding money 423–6
 imperfect commitment 438
 interest rate spread 417
 investment banking 438
 lending collapse 421–2
 liquidity 416
 modern 415–16
 monetary base 419–423
 money multiplier 419–22

money supply 417, 419–23
moral hazard 438
narrow money 419–423
open market operations 436
prudential regulation 437–9
regulation 438–9
required reserve ratio 436
reserve ratio 416–17
reserve requirements 436
retail banking 438
sight deposits 416
time deposits 416

barter economy
 medium of exchange 413–14
 vs monetary economy 415
 money 413–415

Basel III Accord, banking 439

basic prices, gross domestic product (GDP) 353–4

Becker, Gary, economic incentives 34–5

behavioural economics 10, 36

behavioural law 21–2

benefits principle, taxation 327–8

Bernanke, Ben, transmission lag 450

Bertrand model 210
 vs Cournot model 210

beta, asset returns 282–4

bills, financial markets 417–18

bonds, financial markets 418

borrowing, firms' accounts 119

Bretton Woods system, exchange rate regimes 587

BRICs (Brazil, Russia, India and China), economic activity 359–60

broad money, banking 421–423

budget constraint
 budget line 89–96
 consumer choice 89–96

budget deficit 387

budget, government *see* government budget

budget line
 budget constraint 89–96
 chosen bundle 92–4
 price changes 98

budget share 73

budget, structural 395–8

budget surplus (deficit) 320–1

bus transportation, marginal cost 126

business cycles 344, 606–29
 accelerator model of investment 611–13
 ceilings and floors 613

competitiveness 614
credit constraints 617
dynamic stochastic general equilibrium (DSGE) models 618–19
Eurozone 615
international business cycles 619–21
intertemporal substitution 616–617
inventories 613–14
macroeconomic thinking 621–7
multiplier-accelerator model 611–13
output gaps 607
policy implications 617
political business cycles 608
real business cycle theories 615–17
stockbuilding fluctuations 613–14
theories 608–15
trend path of output 607
wages, cyclical behaviour 609–11
business of banking 416
business organization 116–17

C

capital
economic growth 634–5
firms' accounts 118
taxation 509–10
capital accumulation, economic growth 639–40
capital adequacy ratios, banking 437–9
capital assets, price 261
capital controls
 Exchange Rate Mechanism (ERM) 593–4
 fixed exchange rates 564–92
 speculative attacks 589–92
 Tobin tax 568
capital deepening, capital accumulation 639–40
capital flows, gold standard 585–6
capital gain 278
 risk 286
 speculative bubbles 286
capital-labour ratios, comparative advantage 660–1
capital mobility
 fiscal policy 570–1
 monetary policy 570
 open economy macroeconomics 565–6
 perfect capital mobility 554, 565–6
 role 565–6
capital services
 adjustment in the market 259–61

demand 256–7
equilibrium 259–61
industry demand curves 257
long-run adjustment 260
long-run industry supply curve 259
long-run supply 258
long-run supply curve for the economy 258–9
required rental on capital 258
short-run adjustment 260
short-run supply 257–8
supply 257–9

capital stock, unemployment 539
capital widening, capital accumulation 639–40
car industry
 gainers/losers from trade 664–5
 international trade 664–5
Carney, Mark, nominal GDP target 490
cars, income elasticity of demand 75
cartels
 oligopoly 203
 Prisoner's Dilemma 205–6
cash, financial markets 417
cash flow, firms' accounts 118
cash reserves, banking 416
ceilings and floors, business cycles 613
central banks 419, 434–56
 see also Bank of England; banking; European Central Bank (ECB); financial markets
 Eurosystem 434–5
 inflation 479–81
 interest rates 443, 475
 lender of last resort 436–7
 monetary control 442–5
 monetary policy 445–82
 money supply 439–41
 nominal interest rate 475
 quantitative easing (QE) 443–7
 transmission lag 450–1
chief executive officers (CEOs)
 pay 276
 principal-agent problem 276
China, exchange rates 557–8
choices 3–5
 see also consumer choice
 incentives 4–5
chosen bundle, budget line 92–4
cigarettes, inferior goods 104–5
circular flow 346–2

claimant unemployment 523
classical economists
 macroeconomics 367
 New Classical macroeconomics 623–627
classical model
 crowding out 481–2
 macroeconomics 473
classical unemployment 525
clearing system, banking 415–16
climate change economics 309–13
 cost-benefit analysis 312–13
 Kyoto Protocol 311–13
 prices vs quantities 309–10
 science of climate change 311
 Stern Review 312–13
 United States 309–10
closed economy 542
closed-shop agreements, trade unions 241
Coase theorem, property rights, efficiency, externalities 307
coffee, price elasticity of demand (PED) 70
collusion
 vs competition 201–3
 oligopoly 201–4
command economy 14–15
commercial banks, business of banking 416
commitment
 game theory 205–6
 government 337
commodity composition, international trade 655
communism v capitalism 14–15
companies, business organization 117
company shares (equities), financial markets 418
comparative advantage 12–13
 see also absolute advantage
 capital-labour ratios 660–1
 gains from trade 659, 661–5
 international trade 656–65
 losers from trade 663–5
 opportunity cost 12–13
 unit labour requirements 660
comparative static analysis 47–9
 long-run equilibrium 178
 market demand curve 179
 monopolist 182
 short-run equilibrium 178
comparisons, international, macroeconomics 360–1
competition
 banking 441

vs collusion 201–3
economic growth 636–7
potential competition, entry 213–16
competition policy
mergers 211–12
monopoly power 186–7
takeovers 211–12
competitive equilibrium, free markets 295–8
competitive industry, vs monopolist 184–7
competitiveness, business cycles 614
complements
consumer choice 105–7
demand curve 45–6
confidence
aggregate demand 383–4
role 383–4
constant returns to scale 157
consumer choice 84–15
assumptions 85–89
budget constraint 89–96
complements 105–7
cross-price elasticity of demand 102–5
demand decisions 95–13
diminishing marginal rate of substitution 86
diminishing marginal utility 111–12
Giffen goods 102, 105
income changes adjustment 95–7
income effect 98–100
income expansion paths 96–7
indifference curves 87–9
individual demand curve 100–1
inferior goods 101–5
marginal rate of substitution (MRS) 86
marginal utility 111–12
market demand curve 105
measurable utility 111–12
price changes 97–8
single consumer 84–95
substitutes 105–7
substitution effect 98–100
tastes 85–9
tastes and utility 85–6
transfer payments 107–8
transfers in kind 107–8
utility maximization 92–5
utility, measurable 111–12
consumer credit
consumer durables 447–8

transmission mechanism 447–8
consumer durables, consumer credit 447–8
consumer price index (CPI) 24–6
 CPI inflation 515
consumer surplus 50–2
consumption demand
 aggregate demand 368–370
 life-cycle hypothesis 448–9
 monetary policy 447
 personal disposable income 368–9
consumption function 369–70
 marginal propensity to consume 370
 saving function 370
 wealth effect 447
contestable markets
 entry 213–14
 globalization 213–14
contour maps, indifference curves 89
controlling inflation 512–14
convergence hypothesis, economic
 growth 644–7
cost, firms' accounts 117
cost-benefit analysis, climate change economics 312–13
cost of holding money 423
cost of unemployment 537–9
costs
 economic growth 649
 fixed costs 145–6
 long run 153–6
 short run 145–51
 tariffs 666–7
 variable costs 145–6
Cournot model 206–10
 vs Bertrand model 210
 Nash equilibrium 207–10
 Prisoner's Dilemma 208
 reaction function 209–10
CPI *see* consumer price index
crashes
 see also financial crises
 investment crashes 375–6
creating money, banking 416–17
credibility, government 337
credibility and expectations 503–6
credible threats, game theory 206
credit constraints
 aggregate supply 617
 business cycles 617

credit crunch 7
crises, financial *see* financial crises; financial crisis (2007–8) criticisms
economics 34–6
economists 34–6
cross-price elasticity of demand, consumer choice 102–5
cross-section data 22–3
crowding out
 classical model 481–2
 Keynesian model 464, 481, 482
currency area 595–6
currency boards, fixed exchange rates 589–92
currency wars, exchange rates 560–1
current account
 balance of payments 547
 determinants 550–1
 exports 550–1
 imports 551
current transfers, balance of payments 547
cyclical behaviour, wages 609–11
cyclical fluctuations, unemployment 535–7
cyclical output fluctuations, government budget 395–7

D

data 21–2
 see also economic data
deadweight loss, monopoly 186
debt
 fiscal policy 391–3
 Japan 391–3
decision making
 government 334–7
 interdependent decisions, game theory 204–6
 political economy 334–7
defence, tariffs 669
deficits
 fiscal stance 395–8
 inflation 499–500
 inflation-adjusted deficits 397–8
 money 499–500
 national debt 399–403
deflation 511–12
 Ireland 601–3
deflationary gap 379
demand 40–3
 capital services 256–7
 derived demand 222–3
 excess demand 42–3
 income effect 72–5

income elasticity of demand 72–5
inflation 76
mobile apps 48
taxation 78–80

demand curve 43–8
deriving 112–13
individual demand curve 100–1
kinked demand curve 203–4
oligopoly 203–4
perfect competition 171–2
shifts 46–8

demand decisions
consumer choice 95–13
single consumer 84–95

demand-deficient unemployment 525

demand elasticity *see* price elasticity of demand (PED)

demand for factors, long run 222–3

demand for labour
changes 225–6
short run 223–6

demand for money 423–6
interest rates 424–6
motives for holding money 423–6
prices 424–6
real income 424–6

demand management
fiscal policy/monetary policy 470
macroeconomics 368, 470

demand shocks 488
monetary policy 479

depletable resources 635

depreciation, firms' accounts 118

depreciation/appreciation, foreign exchange market 544

Depression, Little/Great Depression 374–5

derivatives, financial markets 418

derived demand 222–3

devaluation
adjustment 572–3
long run 572
medium run 571–2
open economy macroeconomics 571–3
short run 571

diagrams, economic models 31–3

diamond-water paradox, marginal utility 112

diminishing marginal rate of substitution, consumer choice 86

diminishing marginal returns 141–3

diminishing marginal utility
consumer choice 111–12

risk 270–1
direct taxes 322
disagreement areas, macroeconomic thinking 621
discount rate, banking 436
discounting, present values (PV) 253
discouraged workers, unemployment 522, 538
discretion, monetary policy 458
discretionary fiscal policy 398–9
 automatic stabilizers 398–9
discriminating monopoly 187–90
diseconomies of scale 156
disposable income 388–5
 personal disposable income 368–9
distortions
 taxation 301–1
 welfare economics 300–2
diversification
 asset returns correlation 281–2
 portfolio selection 280–2
dividends 278
domestic price of foreign exchange 542
dominant strategy, game theory 205–6
Draghi, Mario, European Central Bank (ECB) 598–600
DSGE models *see* dynamic stochastic general equilibrium models
dumping
 price controls 55
 tariffs 670
dynamic stochastic general equilibrium (DSGE) models, business cycles 618–19

E

earnings
 firms' accounts 120
 UK weekly 221–2
easyJet
 price elasticity of demand (PED) 71
 Stelios Haji-Ioannou 71
ECB *see* European Central Bank
economic data 22–3
 cross-section data 22–3
 time-series data 22–3
economic growth 28, 344–51
 accumulation 638–41
 capital 634–5
 capital accumulation 639–40
 competition 636–7
 convergence hypothesis 644–7
 costs 649
 economies of scale 635–6

endogenous growth 647–9
factor contributions 635–6
human capital 635
innovation 636–7
invention 636–7
labour 635
land 635, 638–9
Malthusian trap 638–9
neoclassical growth theory 640–1
Organization for Economic Cooperation and Development (OECD) 642–7
population 638–9
poverty trap 642–3
production function 634
raw materials 635
research and development 637
saving rates 640–1
Solow residual 647
steady-state path 639–9
technical knowledge 636–7
technical progress 641–2
zero-growth proposal 649

economic history, facts 344–6
economic incentives 36
 Becker, Gary 34–5
economic integration, vs economic sovereignty 333–4
economic models 28–9
 criticisms 35
 diagrams 31–3
 empirical evidence 29–31
 equations 31–3
 lines 31–3
 London Underground 28–4
 scatter diagrams 30, 31
economic profits vs accounting profits 121
economic rent, labour market 236–8
economic sovereignty
 vs economic integration 333–4
 government 333–4
economic variables, measuring changes 28
economics 2
 criticisms 34–6
economies of scale 155–6
 economic growth 635–6
 globalization 159–60
 income distribution 159–60
 internet 156
 involuntary unemployment 241
 technical change 159–60

economists, criticisms 34–6
Edgeworth box
 equity and efficiency 299–300
 welfare economics 299–300
education, higher *see* higher education
effective exchange rates 544–5
efficiency
 Coase theorem 307
 taxation 329
efficiency and equity *see* equity and efficiency
efficiency wages, labour market 242
efficient asset markets
 asset returns 284–7
 efficiency testing 285–6
 speculative bubbles 286
elastic/inelastic, defining 64
elasticity of demand *see* price elasticity of demand (PED)
elasticity of supply 76–80
 point elasticity of supply 77
 summary 77
 taxation 78–80
electricity supply, price controls 54–5
emerging market economies
 international trade 656
 manufactured exports 656
empirical evidence
 economic models 29–31
 hysteresis 539
employment
 see also unemployment
 imports 407–8
endogenous growth, economic growth 647–9
endogenous variables, aggregate demand 371
entrepreneurship, risk 289
entry
 contestable markets 213–14
 entry barriers, supermarkets 216
 entry/exit, firms 173–4
 globalization 213–14
 innocent entry barriers 197, 214
 potential competition 213–16
 strategic entry deterrence 215–216
 strategic moves 215–216
environmental issues, climate change economics 309–13
equations, economic models 31–3
equilibrium
 aggregate demand 373–7
 capital services 259–61

dynamic stochastic general equilibrium (DSGE) models, business cycles 618–19
financial markets 439–42
general equilibrium vs partial equilibrium 297–8
industry labour market equilibrium 233–6
IS-LM model 462–3
market equilibrium 44–2
money market equilibrium 440–1
monopolistic competition 201–2
Nash equilibrium 205
planned saving/planned investment 376–7
tangency equilibrium 201–2
equilibrium income, net exports 404–5
equilibrium inflation 477–82
equilibrium output 407–8
aggregate demand 372–6
equilibrium price 42–3
see also market equilibrium
price elasticity of demand (PED) 66–7
quantity fluctuations 66–7
equilibrium unemployment 525–8
equilibrium with a tariff, international trade 665–6
equity
market failure 302–3
public goods 302–3
taxation 302–3
equity and efficiency
Edgeworth box 299–300
horizontal equity 293–4
vertical equity 293–4
welfare economics 293–300
ERM *see* Exchange Rate Mechanism
EU Single Market, tariffs 671–2
euro 595–603
European Central Bank (ECB) 597–8
monetary sovereignty 596
optimal currency area 595–6
Stability Pact 597, 601
European Central Bank (ECB)
see also central banks
Draghi, Mario 598–600
euro 597–8
Eurosysten 435
monetary control 442
policy 597–600
transmission lag 450–1
Eurosysten, central banks 434–5
Eurozone
antecedents 593–5

business cycles 615
Exchange Rate Mechanism (ERM) 593–4
exchange rates 557
financial markets 459–61
Ireland, deflation 601–3
Maastricht criteria 594
UK/Sterling 594–5
Eurozone/United States, federal fiscal system 600–1
evidence, theories 34
excess demand 42–3
excess supply 43
Exchange Rate Mechanism (ERM) 593–4
 capital controls 593–4
exchange rate regimes 545–605
 adjustable peg 587
 balance of payments 548
 Bretton Woods system 587
 fixed exchange rates 545–7
 floating exchange rates 546, 548, 588
 gold standard 584–7
exchange rates
 see also foreign exchange market
 China 557–8
 currency wars 560–1
 domestic price of foreign exchange 542
 effective exchange rates 544–5
 Eurozone 557
 exchange rate regimes 545–6
 international trade 657–60
 international value of the domestic currency 542
 long-run equilibrium 556–60
 measuring 542–3
 nominal exchange rate 549–50
 par value 571
 purchasing power parity (PPP) 548, 550, 575
 real exchange rate 549–60
 sterling 578–81
exit/entry, firms 173–4
 see also entry
exogenous expectations, macroeconomic thinking 622
exogenous variables, aggregate demand 371
expectations, supply curve 50
expectations and credibility 503–6
 acceleration hypothesis 505–6
expectations and tastes, demand curve 46
expectations formation, macroeconomic thinking 622–3
expected utility of income, risk 273–5
export demand, foreign trade 407

export subsidies, international trade 673–4
exports 352
 see also foreign trade; imports; international trade
 current account 550–1
 foreign trade 403–9
 net exports, equilibrium income 404–5
external balance 555–6
externalities
 Coase theorem 307
 marginal private cost (MPC) vs marginal social cost (MSC) 304–5
 market failure 303–9
 Olympic Games 305–6
 property rights 306–9
extrapolative expectations, macroeconomic thinking 623
Extreme Keynesians, macroeconomic thinking 626, 627

F

factor contributions, economic growth 635–6
factor markets, summary 260–1
factor prices 152–3
factors of production 139–40
 demand for factors, long run 222–3
 fixed factors of production 140
 variable factors 140
fair gambles, risk 269
fallacy of composition, price elasticity of demand (PED) 70
fan charts, CPI inflation 515
farming, price elasticity of demand (PED) 70
FDI *see* foreign direct investment
federal fiscal system, United States/ Eurozone 600–1
final goods 347–8
financial account 551–5
 balance of payments 547
 foreign direct investment (FDI) 552–3
 interest parity 554–5
 perfect capital mobility 554
 speculation 551
financial assets, financial markets 417
financial crises 427–31
balance of payments 590–1
 banking crises, preventing 438–9
 Depression, Little/Great Depression 374–5
 financial crisis (2007–81) 5–6
 financial panic 427
 liquidity crisis 427
 preventing 438–9
 solvency crisis 427
 supply-side effects 617–18

financial crisis (2007-81)
investment crashes 375–6
psychology 286–7
risk 286–7
sub-prime mortgage market 5–31

financial intermediaries, business of
banking 416

financial markets 417–19
see also banking; central banks
bills 417–18
bonds 418
cash 417
company shares (equities) 418
derivatives 418
equilibrium 439–42
Eurozone 459–61
financial assets 417
money market equilibrium 440–1
perpetuities 418
real income 441
real money supply 439–442
securitization 418–19

financial panic 427

firms
entry/exit 173–4
marginal firms 174–5
ownership and control 122–3
principal-agent problem 122–3

firms' accounts 117–21
accounting costs 120–1
accounting profits vs economic profits 121
balance sheet 119–20
borrowing 119
capital 118
cash flow 118
depreciation 118
earnings 120
economic profits vs accounting profits 121
flow accounts 117
inventories 118
net worth 119
opportunity cost 120–1
physical capital 118
retained earnings 120
stock accounts 119–20
supernormal profit 121
unpaid bills 117–18

firms' supply decision

cost, revenue, profit 123–4
perfect competition 172–5
first-best allocation, welfare economics 302
first-mover advantage 212–13
Stackelberg model 212–13
first theorem of welfare economics 298–300
fiscal drag
inflation 509
taxation 509
fiscal federalism, United States/ Eurozone 600–1
fiscal policy 387
see also monetary policy
austerity 391–3
capital mobility 570–1
crowding out in Keynesian model 464
debt 391–3
demand management 470
discretionary fiscal policy 398–9
fixed exchange rates 570–1
floating exchange rates 578
IS schedule 463–4
Japan 391–3
monetary policy/fiscal policy 466–8
policy mix 466–8
unemployment rate 398–9
world economy 591–2
fiscal stability
aggregate demand 401–3
responsibility 401–3
fiscal stance 395–8
deficits 395–8
Fisher hypothesis
inflation 497–8
real interest rate 497–8
fixed capital, investment 452–3
fixed costs 145–6
fixed exchange rates
aggregate demand 573–4
balance of payments 548
capital controls 564–92
currency boards 589–92
exchange rate regimes 545–7
fiscal policy 570–1
macroeconomic policy 570–1
monetary policy 570
open economy macroeconomics 564–70
shocks 569–70
speculative attacks 588–92

sterilized intervention 565
unsterilized intervention 565

fixed factors of production 140

flexible inflation targeting 489–91

flight from cash, hyperinflation 498

floating exchange rates

- balance of payments 548
- exchange rate regimes 546, 548, 588
- fiscal policy 578
- long run 574–577
- managed float 588
- monetary policy 578
- open economy macroeconomics 574–8
- purchasing power parity (PPP) 548
- short run 575–7

floors and ceilings, business cycles 613

flow accounts, firms' accounts 117

flows 250

- firms' accounts 117
- unemployment 521–3

football tickets

- price elasticity of demand (PED) 60–69
- total expenditure 69 foreign assets, real exchange rate 559–60

foreign debt, real exchange rate 559–60

foreign direct investment (FDI) 552–3

foreign exchange market 542–5

- see also* exchange rates
- appreciation/depreciation 544
- effective exchange rates 544–5
- exchanging currencies 543–5

foreign exchange reserves 545–6

foreign trade

- see also* international trade
- export demand 407
- globalization 408–9
- income determination 403–9
- multiplier 406
- open economies 406

forex market *see* foreign exchange market

forward markets

- hedging 287–8
- risk 287–8

free markets 15, 52–5

- competitive equilibrium 295–8

free-riders 307–324

free trade equilibrium, international trade 665, 666

free trading, Pareto efficiency 299–300

frictional unemployment 524

Friedman, Milton
inflation 496
macroeconomics 367
permanent income hypothesis 448
functional distribution of income 263
future taxes effect
government solvency 468–70
Ricardian equivalence 469–70

G

gains from trade, comparative advantage 661–5
game theory 204–6
commitment 205–6
credible threats 206
dominant strategy 205–6
interdependent decisions 204–6
Nash equilibrium 205
Prisoner’s Dilemma 205
repeated games 206
strategy 204–6
GDP *see* gross domestic product
GDP deflator 357
general equilibrium, vs partial equilibrium 297–8
Giffen goods
consumer choice 102, 105
inferior goods 102, 105
Gini coefficient, income inequality 264–5
Glass-Steagall Act, investment banking 438
globalization 159–60
contestable markets 213–14
economies of scale 159–60
entry 213–14
foreign trade 408–9
income distribution 159–60
international trade 408–656
market structure 199–200
multinationals 199–200
technical change 159–60
GNP *see* gross national product
gold standard 584–7
advantage/disadvantage 586–7
balance of payments adjustment 585–6
capital flows 585–6
government
aggregate demand 388–93
commitment 337
credibility 337
decision making 334–7

economic sovereignty 333–4
foreign countries and domestic government 350–4
government production 325
local government 333
log rolling 335–7
market economy 322–7
median voter 335, 336
policy co-ordination 337
political economy 334–7
revelation of preferences 325
government activity, scope 387–8
government bailout, banking 438–9
government budget 393–5
 cyclical output fluctuations 395–7
 investment 393–5
 saving 393–5
government debt 387–8
government policy
 labour market 531–7
 unemployment 531–7
government regulation, supply curve 50
government revenue 320–40
 see also taxation
 % of GDP 321
government solvency
 future taxes effect 468–70
 Ricardian equivalence 469–70
government spending 320–40
 % of GDP 321
 effect on output 389
 historical perspective 320
 taxation 322, 323
gradualist monetarists, macroeconomic thinking 624–627
Great/Little depression 374–5
Greece, output gaps 485–6
gross debt 344
gross domestic product (GDP) 18, 344
 basic prices 353–4
 comprehensive measure 358–60
 economic output 633–4
 GDP deflator 357
government revenue 321
 government spending 321
 happiness, economics of 634
 hidden economy 356–7
 measuring 353–60
 nominal GDP 355–490
 per capita real GDP 357–8

real GDP 355–7
under-reporting 356–7
shadow economy 356–7
gross investment 250
gross national income (GNI) *see* gross national product (GNP)
gross national product (GNP) 347
growth, economic *see* economic growth
growth rate 28

H

happiness, economics of 361
 gross domestic product (GDP) 634
headline inflation 514
health and safety, welfare economics 314–15
hedging
 forward markets 287–8
 risk 287–8
hidden economy, gross domestic product (GDP) 356–7
higher education
 labour market 238
 signalling 277
higher labour input, supply-side economics 631
higher net tax rate 395
history, economic *see* economic history
holding money
 motives for holding money 423–6
 opportunity cost 426
 rise in money holdings 425–6
horizontal equity 293–4
horizontal LM schedule, monetary policy 464
horizontal long-run industry supply curve 177
horsemeat, expectations and tastes 46
hostile takeovers 122–3
hours of work, individual labour supply 227–9
household wealth
 life-cycle hypothesis 448–9
 transmission mechanism 447
housing industry, demand/supply 56
human capital, economic growth 635
hyperinflation 26, 498
 flight from cash 498
hysteresis
 empirical evidence 539
 macroeconomic thinking 622
 policy implications 539
 unemployment 538–9

I

ii schedule

aggregate demand 476

inflation targeting 474–5

IS schedule 476

immigration, labour market 234

imperfect commitment, banking 438

imperfect competition, market failure 302

imperfectly competitive firm 196–7

imports 352

see also exports; foreign trade; international trade

current account 551

employment 407–8

incentives, choices 4–5

income

national income 347–55

permanent income hypothesis 448

real income 441

income changes adjustment, consumer choice 95–7

income determination, foreign trade 403–9

income distribution 9–10

economies of scale 159–60

functional distribution of income 263

Gini coefficient 264–5

globalization 159–60

inequality 264–5

personal income distribution 263–4

technical change 159–60

transfer payments, income redistribution 326

UK 263–5

world population 9–10

income effect

consumer choice 98–100

price changes 98–100

income effect on demand 72–5

income elasticity of demand 72–5

cars 75

inferior goods 73–5

luxury goods 73–5

necessities 73–5

normal goods 73–5

using 75

income expansion paths, consumer choice 96–7

income inequality 264–5

income tax rates 321–2

incomes policy, controlling inflation 512–13

increasing labour productivity, supply side economics 631–2

increasing returns to scale *see* economies of scale

index numbers 24–6
averages 24
indifference curves
 consumer choice 87–9
 contour maps 89
 individual labour supply 228–9
 ordinal utility 112–13
 tastes 87–9
indirect taxes 322
individual attitudes, risk 269–71
individual demand curve, consumer choice 100–1
individual labour supply
 hours of work 227–9
 indifference curves 228–9
 labour market 227–32
 participation rates 229–32
industry demand curve for labour, labour market 226–7
industry demand curves, capital services 257
industry labour market equilibrium 233–6
industry supply curves 175–7
 long-run industry supply curve 176–7
 short-run industry supply curve 175–7
inelastic/elastic, defining 64
inequality
 Gini coefficient 264–5
 income distribution 9–5
infant industries, tariffs 669
inferior goods
 cigarettes 104–5
 consumer choice 101–5
 demand curve 46
 Giffen goods 102, 105
 income elasticity of demand 73–5
inflation 494–519
 adaptation 508–10
 aggregate demand 474–5
 anticipated inflation 508–10
 Asia 513–14
 Bank of England 479–513
 central banks 479–81
 controlling inflation 512–14
 costs 508–11
 CPI inflation 515
 deficits 499–500
 deflation 511–12
 demand 76
 fiscal drag 509
 Fisher hypothesis 497–8

flexible inflation targeting 489–91
Friedman, Milton 496
headline inflation 514
hyperinflation 26, 498
institutional reform 513
interest rates 497–16
menu costs of inflation 509
monetary policy 496
Monetary Policy Committee (MPC) 514–16
money 495–500
oil price shocks 479–81
output 500–8
Phillips curve 500–9
pure inflation 494
redistribution, unexpected inflation 510
shoe leather costs of inflation 509
stagflation 508
targets 515–16
taxation 509–10
uncertain inflation 511
underlying inflation 514
unemployment 500–8
unexpected inflation 510
inflation accounting 510
inflation-adjusted budget 397–8
inflation-adjusted deficits 397–8
inflation illusion 508
inflation rate 25–6
 hyperinflation 26
 UK 495
inflation targeting 474
 flexible inflation targeting 489–91
 ii schedule 474–5
inflation tax 499–500
inflationary gap 379
information, risk 314–15
innocent entry barriers 197, 214
innovation, economic growth 636–7
input 139–40
input costs, supply curve 49
insider-outsider distinction, unemployment 538
insiders, labour market 242
institutional reform
 controlling inflation 513
 Maastricht Treaty 513
insurance, risk 271–7
insurance market
 asymmetric information 276–7

risk 276–7

interdependence, oligopoly 201–4

interdependent decisions, game theory 204–6

interest parity 554–5

interest rate spread, banking 417

interest rates

see also real interest rate

central banks 443, 475

demand for money 424–6

Fisher hypothesis 497–8

inflation 497–16

intermediate targets 446

long-term 449–52

monetary control 443

monetary policy 458

Monetary Policy Committee (MPC) 514–16

nominal interest rate 254, 475

present values (PV) 250–6

real interest rate 254–475

short-term 449–52

transmission mechanism 447–54

intermediate goods 347–8

intermediate targets

interest rates 446

monetary policy 446

internal balance 555–6

international business cycles 619–21

international comparisons, macroeconomics 360–1

international monetary system 584

international trade 653–76

see also foreign trade

absolute advantage 658

agricultural protection 656

car industry 664–5

commodity composition 655

comparative advantage 656–65

emerging market economies 656

equilibrium with a tariff 665–6

EU Single Market 671–2

exchange rates 657–60

export subsidies 673–4

free trade equilibrium 665, 666

globalization 408–656

intra-industry trade 662–3

law of comparative advantage 657, 658

less-developed countries (LDCs) 655

opportunity cost 656–9

quotas 673

raw material prices 656
refrigeration 663–4
steel industry 665
subsidies vs tariffs 672
tariffs 665–74
trade patterns 654–6
trade policy 665
World Trade Organization (WTO) 673
world trade patterns 654–6

international value of the domestic currency 542
internet, economies of scale 156
intertemporal substitution, business cycles 616–617
intra-industry trade, international trade 662–3
invention, economic growth 636–7
inventories 350
 business cycles 613–14
 firms' accounts 118
inventory instrument, transmission mechanism 453
investment 348–50
 see also asset returns
 crashes 375–6
 fixed capital 452–3
 government budget 393–5
 gross investment 250
 net investment 250
 portfolio selection 279–84
 real interest rate 255–6
investment banking 438
 Glass-Steagall Act 438
 Volcker Rule 438
investment demand
 aggregate demand 371
 transmission mechanism 452
investment demand schedule, transmission mechanism 452–3
investment spending, aggregate demand 371
'invisible foot', local government 333
invisible hand 15
involuntary unemployment 526
 economies of scale 241
 labour market 238–9
IOU money 414
iPhone
 price elasticity of demand (PED) 68–9
 total expenditure 68–9
 Ireland, deflation 601–3
IS-LM model 458–65
 equilibrium 462–3
 IS schedule 458–9

LM schedule 461–2
monetary policy 458–65
IS schedule
aggregate demand 476
fiscal policy 463–4
ii schedule 476
IS-LM model 458–9
isoquants 166–9
production techniques 166–9

J

Japan
fiscal policy 391–3
lessons from 391–3

K

Keynes, John Maynard, macroeconomics 367
Keynesian macroeconomic thinking
Extreme Keynesians 626, 627
Moderate Keynesians 625, 626
New Keynesians 625–627
Keynesian model, crowding out 464, 481, 482
Keynesian unemployment 527
kinked demand curve, oligopoly 203–4
Krugman, Paul, balance of payments crises 590–1
Kyoto Protocol, climate change economics 311–13

L

labour, economic growth 635
labour-augmenting technical progress 641–2
labour-capital ratios, comparative advantage 660–1
labour force 227, 344
higher labour input 631
labour market 521
labour market 221–47
cheap foreign labour 670
demand for factors, long run 222–3
demand for labour, short run 223–6
derived demand 222–3
earnings, UK weekly 221–2
economic rent 236–8
efficiency wages 242
government policy 531–7
higher education 238
immigration 234
individual labour supply 227–32
industry demand curve for labour 226–7

industry labour market equilibrium 233–6
insiders 242
involuntary unemployment 238–9
labour force 521
marginal product of labour (MPL) 223–6
marginal revenue product of labour (MRPL) 235–6
marginal value product of labour (MVPL) 223–235
minimum wage agreements 238–9
outsiders 242
participation rates 521
reservation wages 236–7
supply of labour 227–33
tariffs 670
trade unions 240–1
unemployment 521–4
wage adjustment 483
wage behaviour 482–3
wage discrimination 242–4
labour requirements, comparative advantage 660
LAC *see* long-run average cost
Laffer curve, taxation 332
land 249
best addresses 262
economic growth 635, 638–9
Malthusian trap 638–9
population 638–9
rents 261–3
law of comparative advantage
see also comparative advantage
international trade 657, 658
law of diminishing marginal returns 11–12
LDCs *see* less-developed countries
leakage 349
lender of last resort, central banks 436–7
less-developed countries (LDCs), international trade 655
life-cycle hypothesis
consumption demand 448–9
household wealth 448–9
limited liability, business organization 117
lines, economic models 31–416
liquidity crisis 427
Little/Great Depression 374–5
LM schedule
horizontal LM schedule 464
IS-LM model 461–2
modern interpretation 462
monetary expansion 465
monetary policy 464

money demand shocks 465
money market equilibrium 461–2
LMC *see* long-run marginal cost
local government 333
economic principles 333
'invisible foot' 333
Tiebout model 333
log rolling, government decision making 335–7
London Underground, economic models 28–4
long run
costs 153–6
demand for factors 222–3
isoquants 166–9
macroeconomic thinking 623
price elasticity of demand (PED) 71–2
production 151–3
production techniques 166–9
vs short run 140
long-run adjustment, capital services 260
long-run average cost (LAC) 153–1
long run devaluation 572
long-run equilibrium
exchange rates 556–60
macroeconomic thinking 622
real exchange rate 556–60
long-run equilibrium output 477
long run floating exchange rates 574–577
long-run industry supply curve 176–7
capital services 259
vs short-run industry supply curve 176–7
long-run marginal cost (LMC) 153–1
long-run output decision 160–1
long-run Phillips curve 501–9
long-run supply, capital services 258
long-run supply curve 173
long-run supply curve for the economy, capital services 258–9
long-run total cost (LTC) 153–6
long-term interest rates 449–52
losers from trade, comparative advantage 663–5
lotteries, risk 271
LTC *see* long-run total cost
lump-of-labour fallacy, unemployment 526–7
luxury goods
income elasticity of demand 73–5
tariffs 668–9

M

Maastricht criteria, Eurozone 594

Maastricht Treaty, controlling inflation 513
macroeconomic policy, fixed exchange rates 570–1
macroeconomic thinking
 business cycles 621–7
 disagreement areas 621
 exogenous expectations 622
 expectations formation 622–3
 extrapolative expectations 623
 Extreme Keynesians 626, 627
 gradualist monetarists 624–627
 hysteresis 622
 long run 623
 long-run equilibrium 622
 market clearing 621–2
 Moderate Keynesians 625, 627
 New Classical macroeconomics 623–627
 New Keynesians 625–627
 rational expectations 623
 short run 623
macroeconomics 17–18
 see also New Classical
 macroeconomics; open economy macroeconomics
 actual output/potential output 367–8
 classical economists 367
 classical model 473
 demand management 368, 470
 Friedman, Milton 367
 history, brief 367–8
 Keynes, John Maynard 367
 scope 343–4
Malthusian trap, economic growth 638–9
managed float, floating exchange rates 588
marginal cost 124–34
 bus transportation 126
 output choice 128–30
marginal decisions, sunk costs 151
marginal firms 174–5
marginal private cost (MPC), marginal social cost (MSC) 304–5
marginal product 141
marginal product of labour (MPL), labour market 223–6
marginal propensity to consume 370
marginal propensity to import 404–5
marginal propensity to save, multiplier 381
marginal rate of substitution (MRS)
 consumer choice 86
 marginal utility 114–15
 utility function 114–15
marginal revenue 124, 126–34

monopolist 182
output choice 128–30

marginal revenue product of labour (MRPL), labour market 235–6

marginal social benefit 324–5

marginal social cost (MSC), marginal private cost (MPC) 304–5

marginal tax rate 321
unemployment 531–3

marginal utility
consumer choice 111–12
marginal rate of substitution (MRS) 114–15
utility function 114–15
water-diamond paradox 112

marginal value product of capital (MVPK) 256–7

marginal value product of labour (MVPL), labour market 223–235

market clearing, macroeconomic thinking 621–2

market demand curve
comparative static analysis 179
consumer choice 105

market economy, government 322–7

market equilibrium 44–5
see also equilibrium price
consumer surplus 50–2
producer surplus 50–2

market failure
asymmetric information 303
equity 302–3
externalities 303–9
imperfect competition 302
public goods 302–3
taxation 302–3
welfare economics 302–3

market prices 350

market structure 196–220
evidence 198–9
game theory 204–6
globalization 199–200
interdependence 201–4
monopolistic competition 200–2
multinationals 199–200
N-firm concentration ratio 198–9
oligopoly 201–4
PC industry 199–200
reaction function 206–13
reasons for differing structures 197–200

markets 40
see also contestable markets; efficient asset markets; financial markets; forward markets; free markets; missing markets

POW camp 14

role 13–16
measurable utility, consumer choice 111–12
measures of money 422
median voter 335, 336
medium of exchange
 barter economy 413–14
 money 413–14
medium run devaluation 571–2
menu costs of inflation 509
mergers
 see also takeovers
 competition policy 211–12
merit (demerit) goods 327
MES *see* minimum efficient scale
microeconomics 17–343
minimum efficient scale (MES) 158–198
minimum wage agreements, labour market 238–9
mismatch, unemployment 529, 538–9
missing markets, welfare economics 314
mixed economy 16
mobile apps, demand 48 models 21–2
 see also economic models
Moderate Keynesians, macroeconomic thinking 625, 627
monetary base, banking 419–423
monetary control
 central banks 442–5
 European Central Bank (ECB) 442
 interest rates 443
 traditional methods 435–6
monetary economy, vs barter economy 415
monetary expansion, LM schedule 465
monetary fiscal mix 466
monetary instruments, monetary policy 445–6
monetary objectives, trade-offs 489
monetary policy 458–72
 see also fiscal policy
 capital mobility 570
 central banks 445–82
 consumption demand 447
 demand management 470
 demand shocks 479
 fiscal policy/monetary policy 466–8
 fixed exchange rates 570
 floating exchange rates 578
 horizontal LM schedule 464
 inflation 496
 instruments 445–6
 interest rates 458

intermediate targets 446
IS-LM model 458–65
monetary instruments 445–6
monetary target 458
policy mix 466–8
supply shocks 478–82
targets 445–6
Taylor rule 490–1
temporary supply shocks 487–489
transmission mechanism 450–1

Monetary Policy Committee (MPC)
inflation 514–16
interest rates 514–16

monetary sovereignty 596
euro 596

monetary target, monetary policy 458

monetary union 585

money 413–33
see also banking
barter economy 413–415
broad money 421–423
cost of holding money 423
creating money 416–17
deficits 499–500
demand for money 423–6
functions 413–15
holding money 423–6
inflation 495–500
IOU money 414
measures of money 422
medium of exchange 413–14
money supply 417, 419–435
motives for holding money 423–6
narrow money 419–423
quantity theory of money 495–6
store of value 414
token money 414
types of money 414
unit of account 414

money demand
LM schedule 465
shifts 442
shocks 465

money illusion 28, 477

money market equilibrium 440–1
LM schedule 461–2

money multiplier, banking 419–22

money supply 435

balance of payments 564–5
banking 417, 419–23
central banks 439–41
fall in money supply 441
real money supply 439–6

monopolist 180
comparative static analysis 182
vs competitive industry 184–7
marginal revenue 182
price elasticity of demand (PED) 182
price setting 181–2
profit-maximizing output 180–4

monopolistic competition 196–2
equilibrium 201–2
mutual funds 201
product differentiation 200–2
tangency equilibrium 201–2

monopoly 170, 179–93
deadweight loss 186
discriminating monopoly 187–90
natural monopoly 191–198
social cost of monopoly 185–6
supply curve 187–90
technical change 190–1

monopoly power 182
competition policy 186–7
profit maximization 182–4

monopsony power 235–6
wage discrimination 243

moral hazard
asymmetric information 275–6
banking 438
risk 275–6

motives for holding money 423–6
asset motive 424
precautionary motive 424
transactions motive 423–4

MPC *see* marginal private cost; Monetary Policy Committee

MPL *see* marginal product of labour

MRPL *see* marginal revenue product of labour

MSC *see* marginal social cost

multinationals
globalization 199–200
market structure 199–200

multiplier 393
aggregate demand 380–2
bank deposit multiplier 419–22
foreign trade 406

marginal propensity to save 381
money multiplier 419–22
open economies 406
output 380–2
multiplier-accelerator model, business cycles 611–13
 mutual funds
 monopolistic competition 201
 product differentiation 201
MVPK *see* marginal value product of capital
MVPL *see* marginal value product of labour

N

N-firm concentration ratio, market structure 198–9
narrow money, banking 419–423
Nash equilibrium 205–10
 Cournot model 207–10
 game theory 205
 strategy 205
national accounts 347–55
 summarizing 352
national debt 387–403
 deficits 399–403
national income 347–55
 measuring 347–8
 output, measuring 347–8
National Lottery, risk 271
natural monopoly 191–198
 regulation 192–3
 regulatory capture 192
 two-part tariff 192
necessities, income elasticity of demand 73–5
negative externality, taxation 329–331
neoclassical growth theory, saving rates 640–1
net debt 344–6
net exports, equilibrium income 404–5
net investment 250
net taxes, effect on output 388–9
net worth, firms' accounts 119
New Classical macroeconomics, macroeconomic thinking 623–627
New Keynesians, macroeconomic thinking 625–627
nominal exchange rate 549–50
 sterling 578–81
nominal GDP 355–7
 targets 490
nominal interest rate 254, 475
nominal money growth 482
nominal values 26–8
nominal variables 26–8

non-tariff barriers 671
normal goods
 demand curve 46
 income elasticity of demand 73–5
normal profits 173
normative economics 16–35

O

OECD *see* Organization for Economic Cooperation and Development
oil price shocks 7–8
 inflation 479–81
oligopoly 196–7
 cartels 203
 collusion 201–4
 demand curve 203–4
 interdependence 201–4
 kinked demand curve 203–4
 profits from collusion 201–4
Olympic Games, externalities 305–6
open economies 542
 foreign trade 406
 multiplier 406
open economy macroeconomics 564–83
 capital mobility 565–6
 devaluation 571–3
 fixed exchange rates 564–70
 floating exchange rates 574–8
 macroeconomic policy, fixed exchange rates 570–1
open market operations
 Bank of England 436
 banking 436
open source software paradox 325–6
opportunity cost 3–4, 12–13
 absolute advantage 12–13
 comparative advantage 12–13
 firms' accounts 120–1
 holding money 426
 international trade 656–9
optimal currency area 595–6
optimal tariff 668
ordinal utility, indifference curves 112–13
Organization for Economic Cooperation and Development (OECD), economic growth
 642–7
'other things equal' 29, 31, 33–4
output 139–86
 aggregate demand 365–86
 cyclical output fluctuations,
 government budget 395–7

government spending effect 389
inflation 500–8
macroeconomics, actual output/
potential output 367–8
multiplier 380–2
national income, measuring output 347–8
net taxes 388–9
potential output 365–6
trend path of output 607
unemployment 500–535
output choice, marginal cost/marginal revenue 128–30
output decisions
long-run output decision 160–1
short-run output decision 150–1
output gaps 379, 395–6
business cycles 607
Greece 485–6
output is demand-determined 366
outsiders, labour market 242
ownership and control, firms 122–3
P
par value, exchange rates 571
paradox of thrift 381–3
Pareto efficiency
free trading 299–300
perfect competition 295–300
welfare economics 294–300
partial equilibrium, vs general equilibrium 297–8
participation rates
individual labour supply 229–32
labour market 521
Turkey 232
partnerships, business organization 116–17
pay, chief executive officers (CEOs) 276
PC industry, market structure 199–200
PED *see* price elasticity of demand
per capita real GDP 357–8
percentage change 28 perfect capital mobility 554, 565–6
perfect competition 170–9
demand curve 171–2
firms' supply decision 172–5
Pareto efficiency 295–300
welfare economics 295–300
permanent income hypothesis 448
permanent supply shocks 486–8
perpetuities, financial markets 418
personal disposable income 368–9
personal income distribution 263–4

Phillips curve
aggregate supply 502–3
inflation 500–9
long-run Phillips curve 501–9
short-run Phillips curve 501–9
unemployment 500–9
vertical long-run Phillips curve 501–9

physical capital 248–50
firms' accounts 118

planned saving/planned investment
aggregate demand 376–7
equilibrium 376–7

point elasticity of demand 64–5

point elasticity of supply 77

policy co-ordination, government 337

policy implications
business cycles 617
hysteresis 539

policy mix, fiscal policy/monetary policy 466–8

political business cycles 608

political economy 334–7

pooling risk *see* risk pooling

population
economic growth 638–9
land 638–9
Malthusian trap 638–9

portfolio selection 279–84
diversification 280–2
investment 279–84
risk 279–84
risk pooling 281–2
risk-return choice 280

positive economics 16–35

potential competition, entry 213–16

potential output 365–477

potential output/actual output, macroeconomics 367–8

poverty trap, economic growth 642–3

POW camp
markets 14
scarcity of resources 14

PPF *see* production possibility frontier

PPP *see* purchasing power parity

precautionary motive, holding money 424

present values (PV)
discounting 253
interest rates 251–2

price ceilings 52–4

price changes

budget line 98
consumer choice 97–8
income effect 98–100
price elasticity of demand (PED) 67–9
substitution effect 98–100
total expenditure 67–9

price controls 52–5
dumping 55
electricity supply 54–5

price discrimination, discriminating monopoly 187–90

price elasticity of demand (PED) 60–83
calculations 62–3
coffee 70
cross-price elasticity of demand 72, 102–5
determinants 65–6
easyJet 71
equilibrium price 66–7
fallacy of composition 70
farming 70

football tickets 60–69
iPhone 68–9
long run 71–2
measuring 66
monopolist 182
point elasticity of demand 64–5
price changes 67–9
quantity fluctuations 66–9
short run 71–2
Stelios Haji-Ioannou 71
summary 77
total expenditure 67–9
unit elastic 65
using 66–7

price floors 52–5

price of an asset 250

price setting, monopolist 181–2

prices, demand for money 424–6

prices vs quantities, climate change economics 309–10

principal-agent problem 122–3
chief executive officers (CEOs) 276
takeover threat 122–3

principle of targeting, tariffs 668

Prisoner's Dilemma
cartels 205–6
Cournot model 208
game theory 205

private cost of unemployment 537

private goods 322–4

producer surplus 50–2
product differentiation
 monopolistic competition 200–2
 mutual funds 201
production
 long run 151–3
 short run 141–5
production efficiency 12
production function 139–40
 economic growth 634
 technical progress 152
production possibility frontier (PPF) 11–12
production techniques
 isoquants 166–9
 long run 166–9
productivity puzzle 143–4
profit, firms' accounts 117
profit maximization 122–4
 monopoly power 182–4
profit-maximizing output, monopolist 180–4
profits from collusion, oligopoly 201–4
property rights
 Coase theorem 307
 externalities 306–9
prudential regulation, banking 437–9
psychology
 financial crisis (2008) 286–7
 risk 286–7
public goods 322–4
 equity 302–3
 market failure 302–3
 taxation 302–3
purchasing power of money 27
purchasing power parity (PPP)
 exchange rates 548, 550, 575
 floating exchange rates 548
pure inflation 494
PV *see* present values

Q

QE *see* quantitative easing
quality, welfare economics 314–15
quantitative easing (QE)
 see also seigniorage
 central banks 443–7
quantity fluctuations
 equilibrium price 66–7
 price elasticity of demand (PED) 66–9

quantity theory of money 495–6
quotas, international trade 673

R

rational expectations, macroeconomic thinking 623 raw material prices, international trade 656
raw materials, economic growth 635
reaction function 206–13
 Cournot model 209–10
real business cycle theories 615–17
real exchange rate 549–50
 foreign assets 559–60
 foreign debt 559–60
 long-run equilibrium 556–60
 sterling 578–81
real GDP 355–7
real income
 demand for money 424–6
 financial markets 441
real interest rate 254–475
 Fisher hypothesis 497–8
 investment 255–6
 saving 255–6
real money demand 496–7
real money supply
 financial markets 439–442
 money demand, shifts 442
 quantity theory of money 495–6
real values 26–8
real variables 26–8
redistribution, unexpected inflation 510
refrigeration
 gainers/losers from trade 663–4
 international trade 663–4
regulation
 banking 438–9
 natural monopoly 192–3
regulatory capture, natural monopoly 192
renewable resources 635
rent ceiling, Sweden 53 rental rate 250
rents, land 261–3
replacement rate, unemployment 530 repo 435
required rental on capital, capital services 258
required reserve ratio, banking 436
research and development, economic growth 637
reservation wages, labour market 236–7
reserve ratio, banking 416–17
reserve requirements

banking 436
required reserve ratio 436
resource allocation, welfare economics 294–5
resources scarcity *see* scarcity of resources
responsibility, fiscal stability 401–3
retail banking 438
retail price index (RPI) 25–27
retained earnings, firms' accounts 120
returns to scale 155–60
revenue
 firms' accounts 117
 tariffs 669
reverse repo 435
Ricardian equivalence
 future taxes effect 469–70
 government solvency 469–70
risk 269–91
 adverse selection 276–7
 asset returns 277–9
 asymmetric information 275–7
 attitudes, individual 269–71
 capital gain 286
 diminishing marginal utility 270–1
 entrepreneurship 289
 expected utility of income 273–5
 fair gambles 269
 financial crisis (2008) 286–7
 forward markets 287–8
 hedging 287–8
 individual attitudes 269–71
 information 314–15
 insurance 271–7
 insurance market 276–7
 lotteries 271
 missing markets 314
 moral hazard 275–6
 National Lottery 271
 portfolio selection 279–84
 psychology 286–7
 risk-averse 270
 risk-loving 270
 risk-neutral 269
 risk pooling 271–2
 risk pooling, portfolio selection 281–2
 risk/reward 289
 risk sharing 271–2
 speculation 287–8
 speculative bubbles 286

spot markets 287–8
stock market volatility 278–9
total utility of income 273–5
uncertainty 277–9
utility of income 273–5
risk-averse 270
risk-loving 270
risk-neutral 269
risk pooling 271–2
 portfolio selection 281–2
risk-return choice, portfolio selection 280
risk sharing 271–2
Romer, Paul, endogenous growth 647–8
RPI *see* retail price index
rule, monetary policy 458

S

SAFC *see* short-run average fixed cost
SATC *see* short-run average total cost
Satisfaction with Life index 361
SAVC *see* short-run average variable cost
saving 348–50
 government budget 393–5
 real interest rate 255–6
 vs savings 384
saving function, consumption function 370
saving rates
 aggregate demand 382–3
 neoclassical growth theory 640–1
scale economies *see* economies of scale
scarcity of resources 2, 11–13
 POW camp 14
scatter diagrams, economic models 30, 31
school policy, general equilibrium vs partial equilibrium 297–8
second-best allocation, welfare economics 302
second theorem of welfare economics 298–300
securitization, financial markets 418–19
seigniorage 499
 see also quantitative easing (QE)
self-fulfilling prophecy 504
services, output 4–5
shadow economy, gross domestic product (GDP) 356–7
shocks
 see also demand shocks; oil price shocks; supply shocks
 fixed exchange rates 569–70
shoe leather costs of inflation 509
short run
 costs 145–51

demand for labour 223–6
vs long run 140
macroeconomic thinking 623
price elasticity of demand (PED) 71–2
production 141–5

short-run adjustment, capital services 260

short-run aggregate supply 483–4

short-run average fixed cost (SAFC) 147–51

short-run average total cost (SATC) 147–161

short-run supply curve 172–3

short-run average variable cost (SAVC) 147–51

short run devaluation 571

short-run equilibrium, comparative static analysis 178

short-run equilibrium output, aggregate demand 372

short run floating exchange rates 575–7

short-run industry supply curve 175–7

 vs long-run industry supply curve 176–7

short-run marginal cost (SMC) 146–7

 short-run supply curve 172–3

short-run output decision 150–1

short-run Phillips curve, Phillips curve 501–9

short-run supply, capital services 257–8

short-run supply curve 172–484

 short-run average total cost (SATC) 172–3

 short-run marginal cost (SMC) 172–3

short-term interest rates 449–52

shutdown price 175

sight deposits, banking 416

signalling, higher education 277

Single Market, EU, tariffs 671–2

SMC *see* short-run marginal cost

smoking ban, taxation 329–30

social cost of monopoly 185–6

social cost of unemployment 538–9

sole traders, business organization 116

Solow residual, economic growth 647

solvency crisis 427

speculation 551

 risk 287–8

speculative attacks

 capital controls 589–92

 fixed exchange rates 588–92

speculative bubbles

 capital gain 286

 efficient asset markets 286

spot markets, risk 287–8

Stability Pact, euro 597, 601

stabilization policy 387

automatic stabilizers 398–9
Stackelberg model 212–13
 first-mover advantage 212–13
stagflation 508
 inflation 508
standardized unemployment 523–4
standards, imposing 315
steady-state path
 capital accumulation 639–40
 economic growth 639–9
 endogenous growth 647–9
steel industry
 gainers/losers from trade 665
 international trade 665
Stelios Haji-Ioannou
 easyJet 71
 price elasticity of demand (PED) 71
sterilized intervention, fixed exchange rates 565
sterling
 exchange rates 578–81
 nominal exchange rate 578–81
 real exchange rate 578–81
Stern Review, climate change economics 312–13
stock accounts, firms' accounts 119–20
stock market volatility, risk 278–9
stockbuilding fluctuations, business cycles 613–14
stocks 250
 see also inventories
 firms' accounts 117
 unemployment 521–3
store of value, money 414
strategic entry deterrence 215–216
strategic moves, entry 215–216
strategic trade policy, tariffs 669–70
strategy
 dominant strategy 205–6
 game theory 204–6
 Nash equilibrium 205
structural budget 395–8
structural unemployment 524–5
sub-prime mortgage market, financial crisis (2007–81) 5–31
subsidies vs tariffs, international trade 672
substitutes
 consumer choice 105–7
 demand curve 45–6
substitution effect
 consumer choice 98–100
 price changes 98–100

sunk costs, marginal decisions 151
supermarkets, entry barriers 216
supernormal profit, firms' accounts 121
supply 40–3
 see also elasticity of supply
 capital services 257–9
 excess supply 43
 theory 139–40
supply curve 43–50
 expectations 50
 government regulation 50
 industry supply curves 175–7
 input costs 49
 monopoly 187–90
 shifts 50
 taxation 78–80
 technology 49
supply decision, firms'
 competitive firms 175
 cost, revenue, profit 123–4
 perfect competition 172–5
supply of labour
 individual labour supply 227–32
 industry 232–3
 labour market 227–33
supply shocks 486–8
 monetary policy 478–82
 oil price shocks 7–8, 479–81
 permanent supply shocks 486–8
 temporary supply shocks 487–489
supply-side economics 630–2
 higher labour input 631
 increasing labour productivity 631–2
 taxation 332
 unemployment 529–33
supply-side effects, financial crises 617–18
supply-side factors, unemployment 529–33
Sweden, rent ceiling 53

T

takeovers
 see also mergers
 competition policy 211–12
 hostile takeovers 122–3
 takeover threat, principal-agent problem 122–3
tangency equilibrium, monopolistic competition 201–2
tangible wealth 249
targets

flexible inflation targeting 489–91
inflation 515–16
intermediate targets 446
monetary policy 445–6
monetary target, monetary policy 458
nominal GDP 490

tariffs

arguments for/against 668–72
benefits 666–7
cheap foreign labour 670
costs 666–7
defence 669
dumping 670
equilibrium with a tariff 665–6
EU Single Market 671–2
infant industries 669
international trade 665–74
labour market 670
levels 672–3
luxury goods 668–9
non-arguments 670
non-tariff barriers 671
optimal tariff 668
principle of targeting 668
reasons for 671–2
revenue 669
strategic trade policy 669–70
vs subsidies 672
trade policy 665, 669–70
way of life 668–9
World Trade Organization (WTO)

673

tastes

assumptions 85–89
consumer choice 85–9
indifference curves 87–9
tastes and expectations, demand curve 46

tastes and utility, consumer choice 85–6

tax incidence 328–9

tax wedge, unemployment 531

taxation

see also government revenue

benefits principle 327–8

capital 509–10

demand 78–80

distortions 301–1

effect, specific tax 79–329

elasticity of supply 78–3

fair 327–8
fiscal drag 509
future taxes effect 468–70
government solvency 468–70
government spending 322, 323
incidence of tax 79
income tax rates 321–2
inflation 509–10
inflation tax 499–500
Laffer curve 332
marginal tax rate 321, 531–3
market failure 302–3
negative externality 329–331
net taxes 388–9
principles 327–31
public goods 302–3
Ricardian equivalence 469–70
smoking ban 329–30
supply-side economics 332
tax incidence 328–9
tax wedge 531
Tobin tax 568
types of taxes 327
waste 329

Taylor rule, monetary policy 490–1

technical change
economies of scale 159–60
globalization 159–60
income distribution 159–60
monopoly 190–1

technical knowledge, economic growth 636–7

technical progress
economic growth 641–2
labour-augmenting technical progress 641–2
production function 152

technically efficient 139–40

technology, supply curve 49

temporary supply shocks 487–489
monetary policy 487–489

theories 21– 2
evidence 34

Tiebout model, local government 333

time, missing markets 314

time deposits, banking 416

time-series data 22–568

token money 414

total expenditure
football tickets 69 iPhone 68–9

price changes 67–9
 price elasticity of demand (PED) 67–9
total utility of income, risk 273–5
trade balance 404–5
trade deficit 404–5
trade, foreign *see* foreign trade
trade-offs, monetary objectives 489
trade patterns, international trade 654–6
trade policy
 international trade 665
 strategic trade policy 669–70
 tariffs 665, 669–70
trade surplus 404–5
trade unions
 closed-shop agreements 241
 labour market 240–1
 unemployment, trade union power 530–1
transactions motive, holding money 423–4
transfer payments 322, 350
 consumer choice 107–8
 income redistribution 326
transfers in kind, consumer choice 107–8
transmission lag
 central banks 450–1
 transmission mechanism 450–1
transmission mechanism 447–54
 consumer credit 447–8
 household wealth 447
 interest rates 447–54
 inventory instrument 453
 investment demand 452
 investment demand schedule 452–3
 monetary policy 450–1
 transmission lag 450–1
 wealth effect 447
trend path of output 607
the ‘tube’ *see* London Underground
Turkey, participation rates, individual labour supply 232
two-part tariff, natural monopoly 192

U

UK inflation rate 495
UK/Sterling, Eurozone 594–5
UK unemployment rate 520–528
uncertain inflation 511
uncertainty
 asset returns 277–9
 risk 277–9

underlying inflation 514
unemployment 520–41
 analysing 524–8
 capital stock 539
 changes, explaining 528–34
 claimant unemployment 523
 classical unemployment 525
 composition 524
 cost 537–9
 cyclical fluctuations 535–7
 demand-deficient unemployment 525
 discouraged workers 522, 538
 duration 522–3
 equilibrium unemployment 525–8
 flows 521–3
 frictional unemployment 524
 government policy 531–7
 hysteresis 538–9
 inflation 500–8
 insider-outsider distinction 538
 involuntary unemployment 238–526
 Keynesian unemployment 527
 labour market 521–4
 lump-of-labour fallacy 526–7
 marginal tax rate 531–3
 measuring 523–4
 mismatch 529, 538–9
 output 500–535
 Phillips curve 500–9
 potential extent 535–7
 private cost of unemployment 537
 replacement rate 530
 social cost of unemployment 538–9
 stagflation 508
 standardized unemployment 523–4
 unemployment#8212; *Cont.*
 stocks 521–3
 structural unemployment 524–5
 supply-side economics 529–33
 supply-side factors 529–33
 tax wedge 531
 trade union power 530–1
 types 524–5
 voluntary unemployment 526
 wages, cyclical behaviour 609–11
 unemployment rate 18, 344–6
 defining 521
 fiscal policy 398–9

UK unemployment rate 520–528
unexpected inflation 510
unit elastic, price elasticity of demand (PED) 65
unit labour requirements, comparative advantage 660
unit of account, money 414
United States, climate change economics 309–10
United States/Eurozone, federal fiscal system 600–1
unpaid bills, firms' accounts 117–18
unsterilized intervention, fixed exchange rates 565
utility function
 marginal rate of substitution (MRS) 114–15
 marginal utility 114–15
utility maximization, consumer choice 92–5
utility of income, risk 273–5

V

value added 347–8
variable costs 145–6
variable factors 140
variables
 endogenous variables 371
 exogenous variables 371
vertical equity 293–4
vertical long-run Phillips curve 501–9
Volcker Rule, investment banking 438
voluntary unemployment 526

W

wage adjustment, labour market 483
wage behaviour, labour market 482–3
wage discrimination, labour market 242–4
wage gap, men/women 243–4
wages
 cyclical behaviour 609–11
 efficiency wages 242
 minimum wage agreements 238–9
 reservation wages 236–7
waste, taxation 329
water-diamond paradox, marginal utility 112
way of life, tariffs 668–9
wealth effect
 consumption function 447
 transmission mechanism 447
welfare economics 293–319
climate change economics 309–13
distortions 300–2
Edgeworth box 299–300

equity and efficiency 293–300
first-best allocation 302
first theorem of welfare economics 298–300
health and safety 314–15
market failure 302–3
missing markets 314
Pareto efficiency 294–300
perfect competition 295–300
quality 314–15
resource allocation 294–5
second-best allocation 302
second theorem of welfare economics 298–300
world economy, fiscal policy 591–2
world population, income distribution 9–10
World Trade Organization (WTO)
 international trade 673
 tariffs 673
world trade patterns, international trade 654–6
WTO *see* World Trade Organization

Z

zero-growth proposal, economic growth 649