AQA Approved

AQA A-level

For A-level
Years 1 and 2

Economics

Fifth Edition

Ray Powell
James Powell
David Horner
Steve Stoddard



Boost

HODDER EDUCATION
LEARN MORE

This page intentionally left blank

AQA
A-level

For A-level
Years 1 and 2

Economics

Fifth Edition

Ray Powell
James Powell
David Horner
Steve Stoddard

Approval message from AQA

This textbook has been approved by AQA for use with our qualification. This means that we have checked that it broadly covers the specification and we are satisfied with the overall quality. Full details of our approval process can be found on our website.

We approve textbooks because we know how important it is for teachers and students to have the right resources to support their teaching and learning. However, the publisher is ultimately responsible for the editorial control and quality of this book.

Please note that when teaching the ***AQA A-level Economics*** course, you must refer to AQA's specification as your definitive source of information. While this book has been written to match the specification, it does not provide complete coverage of every aspect of the course.

A wide range of other useful resources can be found on the relevant subject pages of our website: www.aqa.org.uk.

Boost

HODDER EDUCATION
AN HACHETTE UK COMPANY

Every effort has been made to trace all copyright holders, but if any have been inadvertently overlooked, the Publishers will be pleased to make the necessary arrangements at the first opportunity.

Although every effort has been made to ensure that website addresses are correct at time of going to press, Hodder Education cannot be held responsible for the content of any website mentioned in this book. It is sometimes possible to find a relocated web page by typing in the address of the home page for a website in the URL window of your browser.

Hachette UK's policy is to use papers that are natural, renewable and recyclable products and made from wood grown in well-managed forests and other controlled sources. The logging and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

Orders: please contact Hachette UK Distribution, Hely Hutchinson Centre, Milton Road, Didcot, Oxfordshire, OX11 7HH. Telephone: +44 (0)1235 827827. Email education@hachette.co.uk. Lines are open from 9 a.m. to 5 p.m., Monday to Friday. You can also order through our website: www.hoddereducation.co.uk

ISBN: 978 1 3983 7519 2

© 2023 Ray Powell, James Powell, David Horner and Steve Stoddard

First published in 2023 by

Hodder Education
An Hachette UK Company
Carmelite House
50 Victoria Embankment
London EC4Y 0DZ

www.hoddereducation.co.uk

Impression number 10 9 8 7 6 5 4 3 2 1
Year 2027 2026 2025 2024 2023

All rights reserved. Apart from any use permitted under UK copyright law, no part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or held within any information storage and retrieval system, without permission in writing from the publisher or under licence from the Copyright Licensing Agency Limited. Further details of such licences (for reprographic reproduction) may be obtained from the Copyright Licensing Agency Limited, www.cla.co.uk

Cover photo: William/stock.adobe.com. Other photographs: **p1** Spotmatik/stock.adobe.com; **p6** Kristian Buus/Getty Images; **p13** Ashley Cooper pics/Alamy Stock Photo; **p20** Roland Nagy/Alamy Stock Photo; **p38** UPI/Alamy Stock Photo; **p41** See Li/Alamy Stock Photo (t) and Martyn Evans/Alamy Stock Photo (b); **p49** exclusive-design/stock.adobe.com; **p61** Andreas/stock.adobe.com; **p78** Roger Bamber/Alamy Stock Photo; **p84** Monkey Business/stock.adobe.com; **p90** aitormmfoto/stock.adobe.com; **p97** Lou Linwei/Alamy Stock Photo; **p111** Robert Convery/Alamy Stock Photo; **p115** Stephen Chung/Alamy Stock Photo; **p124** Tetiana Vitsenko/Alamy Stock Photo; **p129** GL Archive/Alamy Stock Photo; **p144** Steve Stock/Alamy Stock Photo; **p152** LStockStudio/stock.adobe.com; **p157** Kevin Clark/Alamy Stock Photo; **p170** Jeff Gilbert/Alamy Stock Photo; **p179** Greatstock/Alamy Stock Photo; **p188** Monkey Business/stock.adobe.com; **p194** Bloomberg/Getty Images; **p201** david pearson/Alamy Stock Photo; **p209** Graham Mulrooney/Alamy Stock Photo; **p213** Daryl Mulvihill/Alamy Stock Photo; **p234** cheekylorns/stock.adobe.com; **p239** Denver Post/Getty Images; **p243** Trinity Mirror/Mirrorpix/Alamy Stock Photo; **p255** Malcolm Park editorial/Alamy Stock Photo; **p259** Advertising Archives; **p264** IanDagnall Computing/Alamy Stock Photo; **p270** Benny Marty/Alamy Stock Photo; **p287** Francesco Scatena/stock.adobe.com; **p297** leungchopan/stock.adobe.com; **p300** IanDagnall Computing/Alamy Stock Photo; **p311** World History Archive/Alamy Stock Photo; **p317** Islandstock/Alamy Stock Photo; **p326** Fir Mamat/Alamy Stock Photo; **p335** The History Collection/Alamy Stock Photo; **p348** Poppy/Alamy Stock Photo; **p352** justaa/stock.adobe.com; **p373** Danita Delimont Creative/Alamy Stock Photo; **p386** Sipa US/Alamy Stock Photo; **p402** Bloomberg/Getty Images; **p424** Prisma by Dukas Presseagentur GmbH/Alamy Stock Photo; **p441** Islandstock/Alamy Stock Photo; **p461** Medicimage Education/Alamy Stock Photo; **p467** Imageplotter/Alamy Stock Photo; **p480** Zefrog/Alamy Stock Photo; p494 Maureen McLean/Alamy Stock Photo; **p512** Peter Charlesworth/Getty Images; **p530** White House Photo/Alamy Stock Photo; p548 DANIEL SORABJI/Getty Images; **p567** JADE GAO/Getty Images; **p584** Tony Watson/Alamy Stock Photo; **p594** sborisov/stock.adobe.com (t) and REUTERS/Alamy Stock Photo (b); **p597** dpa picture alliance archive/Alamy Stock Photo.

Typeset by Integra Software Services Pvt. Ltd., Pondicherry, India

Printed in Bosnia & Herzegovina

A catalogue record for this title is available from the British Library.

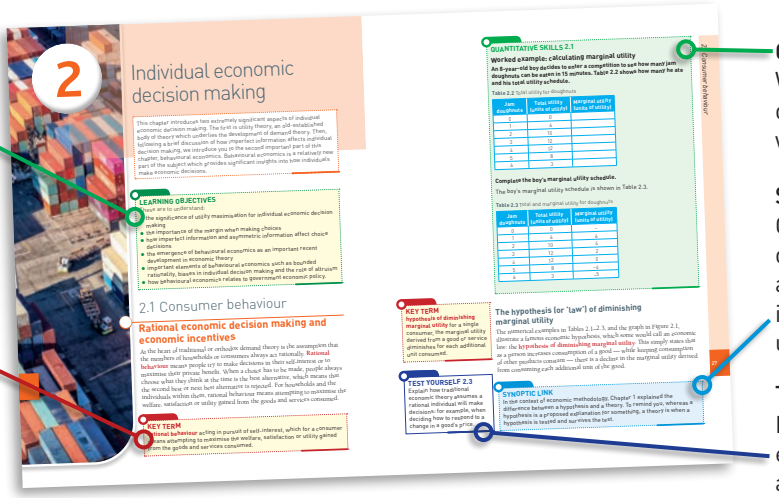


Get the most from this book

This textbook has been tailored explicitly to cover the content of the AQA specification for the A-level course. The book is divided into two parts, one covering microeconomics and the other macroeconomics.

The text provides the foundation for studying AQA economics, but you will no doubt wish to keep up to date by referring to additional topical sources of information about economic events. This can be done by reading the serious newspapers, visiting key sites on the internet and reading such magazines as *Economics Review*.

Special features



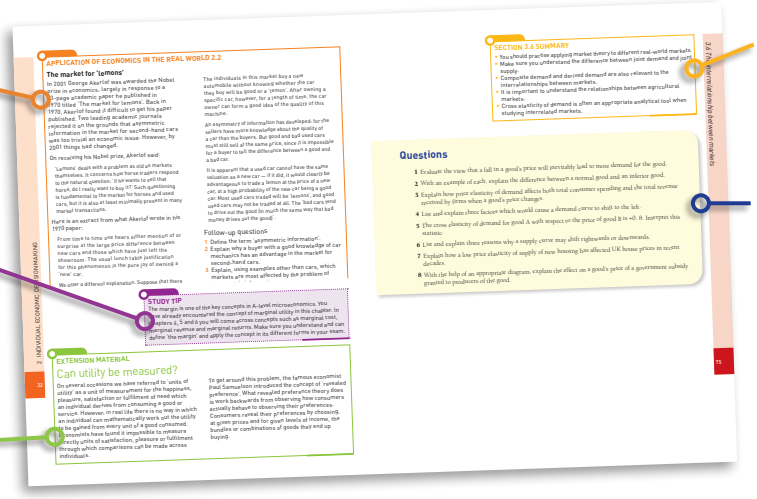
Learning objectives
A statement of the intended learning objectives for each chapter.

Key terms
Clear, concise definitions of essential key terms where they first appear and as a list at the end of each section.

Quantitative skills
Worked examples of the quantitative skills that you will need to develop.

Synoptic links
Connections between different areas of economics are highlighted to help improve your overall understanding of the subject.

Test yourself
Exercises to provide active engagement with economic analysis.



Application of economics in the real world
Case studies to show economic concepts applied to real-world situations.

Study tips
Short pieces of advice to help you present your ideas effectively and avoid potential pitfalls.

Extension material
Extension points to stretch your understanding.

Summaries
Bulleted summaries of each topic that can be used as a revision tool.

Questions
Revision questions are provided at the end of each section to help you check your knowledge and understanding of the topics you have covered within each chapter.

Practice questions
There is a section at the back containing exam-style questions for you to practise.

Answers are online at: www.hoddereducation.co.uk/aqa-economics-alevel

Contents

Part 1 Microeconomics: individuals, firms, markets and market failure

1 Economic methodology and the economic problem

1.1 Economic methodology 2
1.2 The nature and purpose of economic activity 7
1.3 Economic resources 11
1.4 Scarcity, choice and the allocation of resources 13
1.5 Production possibility diagrams 17

2 Individual economic decision making

2.1 Consumer behaviour 24
2.2 Imperfect information 31
2.3 Aspects of behavioural economic theory 33
2.4 Behavioural economics and economic policy 40

3 Price determination in a competitive market

3.1 The determinants of demand 45
3.2 Price, income and cross elasticities of demand 50
3.3 The supply of goods and services 57
3.4 Price elasticity of supply 60
3.5 The determination of equilibrium market prices 64
3.6 The interrelationship between markets 69

4 Production, costs and revenue

4.1 Production 76
4.2 Specialisation, division of labour and exchange 82
4.3 The law of diminishing returns and returns to scale 85
4.4 Costs of production 90
4.5 Economies and diseconomies of scale 94
4.6 Marginal, average and total revenue 103
4.7 Profit 108
4.8 Technological change 112

5 Perfect competition, imperfectly competitive markets and monopoly

5.1 Market structures 120
5.2 The objectives of firms 124
5.3 Perfect competition 129
5.4 Monopoly and monopoly power 135
5.5 The advantages and disadvantages of monopoly 137
5.6 Monopolistic competition 140
5.7 Oligopoly 145
5.8 Price discrimination 153

5.9 The dynamics of competition and competitive market processes ... 155
5.10 Contestable and non-contestable markets ... 156
5.11 Market structure, static efficiency, dynamic efficiency and resource allocation ... 159
5.12 Consumer and producer surplus ... 159

6 The labour market

6.1 The demand for labour ... 164
6.2 Influences on the supply of labour to different labour markets ... 171
6.3 Perfectly competitive labour markets ... 175
6.4 Imperfectly competitive labour markets ... 178
6.5 The influence of trade unions in determining wages and levels of employment ... 182
6.6 The national minimum wage ... 186
6.7 Discrimination in the labour market ... 191

7 The distribution of income and wealth: poverty and inequality

7.1 The distribution of income and wealth ... 197
7.2 The problem of poverty ... 210
7.3 Government policies to alleviate poverty and to influence the distribution of income and wealth ... 214

8 The market mechanism, market failure and government intervention in markets

8.1 How markets and prices allocate resources ... 225
8.2 The meaning of market failure ... 228
8.3 Private goods, public goods and quasi-public goods ... 229
8.4 Positive and negative externalities in production and consumption ... 235
8.5 Merit and demerit goods ... 252
8.6 Market imperfections ... 262
8.7 Competition policy ... 265
8.8 Public ownership, privatisation, regulation and deregulation of markets ... 271
8.9 Government intervention in markets ... 278
8.10 Government failure ... 285

Microeconomics key terms ... 290

Part 2 Macroeconomics: the national and international economy

9 The measurement of macroeconomic performance

9.1 The objectives of government economic policy ... 298
9.2 Macroeconomic indicators ... 312
9.3 The uses of index numbers ... 314
9.4 The uses of national income data ... 321

10 How the macroeconomy works: the circular flow of income, aggregate demand/aggregate supply analysis and related concepts

10.1 The circular flow of income 328
10.2 Aggregate demand and aggregate supply analysis 336
10.3 The determinants of aggregate demand 343
10.4 Aggregate demand and the level of economic activity 352
10.5 Determinants of short-run aggregate supply 358
10.6 Determinants of long-run aggregate supply 359

11 Economic performance

11.1 Economic growth and the economic cycle 366
11.2 Employment and unemployment 382
11.3 Inflation and deflation 395
11.4 Conflicts between macroeconomic policy objectives 411

12 Financial markets and monetary policy

12.1 The structure of financial markets and financial assets 418
12.2 Commercial banks and investment banks 432
12.3 Central banks and monetary policy 442
12.4 The regulation of the financial system 456

13 Fiscal policy and supply-side policies

13.1 Fiscal policy 464
13.2 Supply-side policies 496

14 The international economy: globalisation and international trade

14.1 Globalisation 510
14.2 Trade 520
14.3 The balance of payments 541

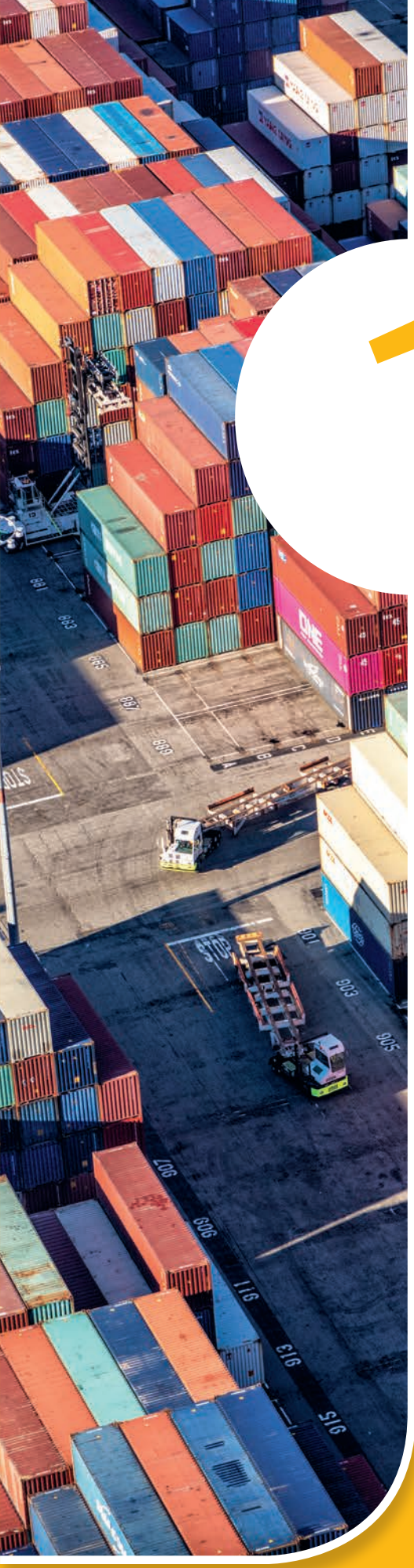
15 Exchange rate systems and economic growth and development

15.1 Exchange rate systems 570
15.2 Economic growth and development 590

Macroeconomics key terms 605

Practice questions 613

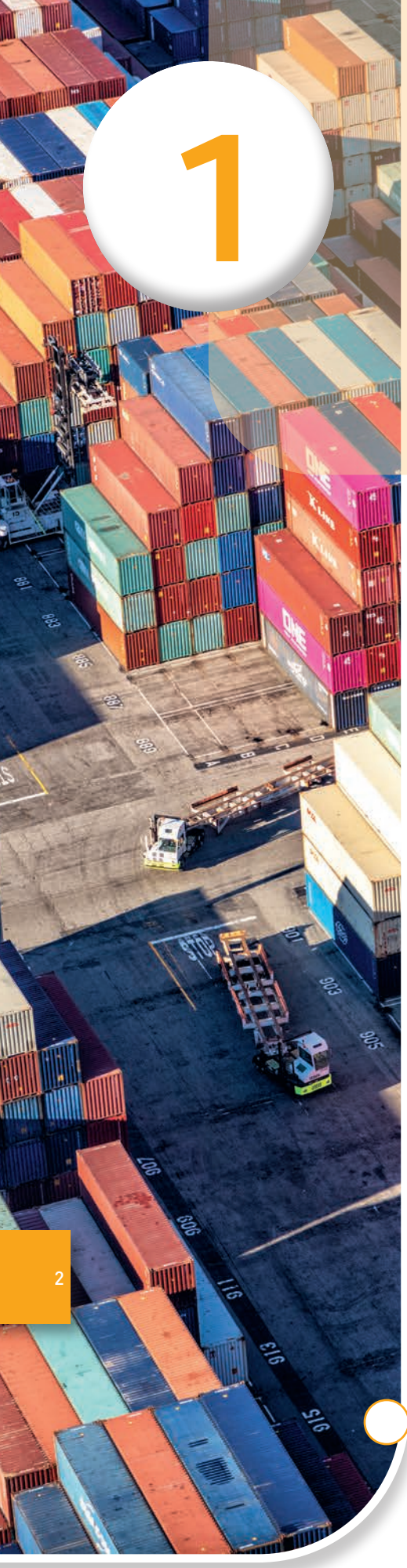
Index 625



1

Microeconomics

Individuals, firms, markets and market failure



1

Economic methodology and the economic problem

Economics is the study of choice and decision making in a world with limited resources. It tries to explain the economic behaviour of both individuals and groups of people, and the economic relationships between individuals and groups. Related to the noun 'economics' is the verb 'to economise'. In large part, economics is the study of economising — the study of how people make choices about what to produce, how to produce and for whom to produce, in a world in which most resources are limited or scarce. How best can people make decisions about how scarce resources should be allocated among competing uses, so as to improve and maximise human happiness and welfare? This is the economic problem, which is the main focus of this introductory chapter.

LEARNING OBJECTIVES

These are to understand:

- economics as a social science
- similarities and differences in methodologies from natural and other sciences
- the difference between positive and normative statements
- how value judgements influence economic decision making and policy
- that the central purpose of economic activity is the production of goods and services to satisfy needs and wants
- the key economic decisions: what to produce, how to produce and who is to benefit from the goods and services produced
- the economists' classification of economic resources into land, labour, capital and enterprise, which are known as the factors of production
- that the environment is a scarce resource
- that the fundamental economic problem is scarcity resulting from limited resources and unlimited wants
- scarcity means that choices have to be made about how resources are allocated
- how production possibility diagrams may be used to illustrate resource allocation, opportunity cost, trade-offs, unemployment, economic growth and efficiency.

1.1 Economic methodology

Economics as a social science

When answering the question 'What is economics?', a good place to start is the fact that economics is a social science. Social science is the branch of science that studies society and the relationships of individuals within a society.

Psychology, sociology and political science are also social sciences, as are important elements of history and geography.

Psychology studies the behaviour and mental processes of an *individual*. Sociology studies the *social* relationships between people in the context of *society*. By contrast, economics, as the name suggests, studies the *economic behaviour* of both individuals and groups of people, and the *economic relationships* between individuals and groups.

Let us give you two examples of what we mean. Our first example (about *individual behaviour*) is from an important part of economics known as demand theory, which is covered in Chapter 3. The theory addresses consumer behaviour, or how we behave when we go shopping. Why, for example, do people generally buy more strawberries as the price of strawberries falls?

Our second example introduces an important *economic relationship*. Having explained demand, we must go a stage further and look at how consumers interact with firms or producers. Firms supply and sell the goods that consumers buy, and economists call the 'place' in which goods are bought and sold a market. Indeed, before you started this economics course, you may well have heard the phrase 'supply and demand' and thought that is what economics is about. Well, in large measure that is true, particularly in the early chapters of this book, which cover Unit 1 of the specification.

TEST YOURSELF 1.1
What is economics the study of?

Economics and scientific methodology

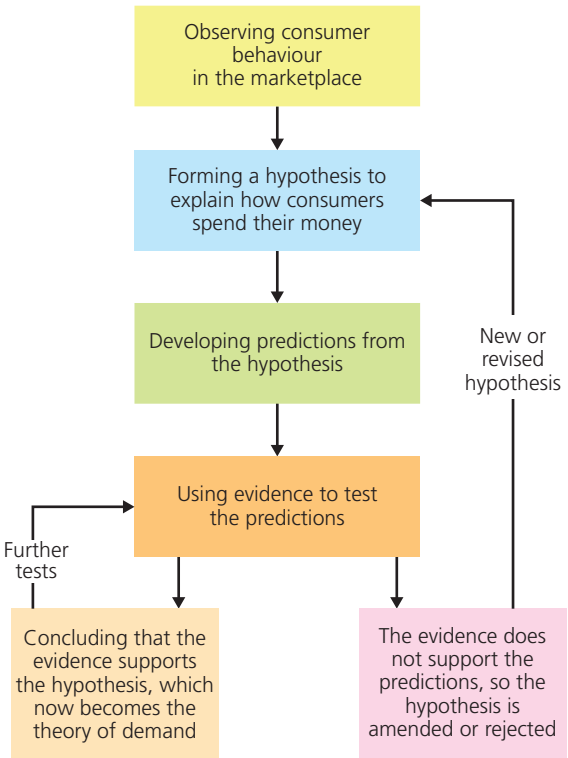


Figure 1.1 An example of how scientific methodology may be used in economics

The essentials of scientific methodology, in the context of the demand theory we will look at in Chapter 3, are shown in the flowchart in Figure 1.1. Scientists start off by observing some aspect of the universe (in the natural sciences), or some aspect of human behaviour, in the case of the social sciences. In the case of demand theory, the starting point — shown in the uppermost box of Figure 1.1 — is observations of how individual consumers react to changes in the prices of the goods and services they buy. Demand theory then develops from making a tentative description, known as a *hypothesis*, of what has been observed. Hypothesis construction is depicted in the second box from the top in the flowchart. In the third box, predictions about human behaviour are deduced from the hypothesis, such as that an individual will always respond to a lower price by demanding more of the good in question.

This prediction is then tested against collected evidence about how individuals behave in the marketplace (the fourth box from the top). At this stage, the hypothesis becomes a *theory*. (The difference between the two is that, whereas a hypothesis is a proposed explanation for something, a theory is when a hypothesis is tested and survives the test.)

At this stage, we are in the bottom left-hand box of Figure 1.1. However, this does not mean that the theory is true in all circumstances. All it says is that the hypothesis has survived the test or tests to which it has been exposed. It might not survive stronger tests, which may not yet have been devised. Scientific method is based on the possibility of *falsification* or *refutation* of a hypothesis.

If a hypothesis fails to survive the tests to which it is exposed, one of two things can happen. The bottom right-hand box of Figure 1.1 shows the first possibility: outright rejection of the hypothesis. For example, a hypothesis that consumers *always* respond to price cuts by demanding *less* would surely be rejected (as would the hypothesis that consumers *always* respond to a price cut by demanding more of a good). The other possibility, which often occurs in economics, is that the hypothesis is changed, usually by diluting it, so as to make it less deterministic. In the case of demand theory, as the extension material on page 49 in Chapter 3 explains, diluting it means that demand theory predicts that, *in most but not all cases*, consumers respond to price cuts by demanding more of a good. Diluted in this way, the demand hypothesis survives the tests to which it is exposed, and becomes 'the first "law" of demand'.

TEST YOURSELF 1.2
Give another example of how a theory can be developed from a hypothesis in economics.

Social sciences and natural sciences

Economic theories often survive only through allowing a significant number of exceptions to their central predictions, which, according to critics, turns the theories into little more than generalisations with low scientific application.

To quote the economist Ha-Joon Chang, economics is not a science like biology is a science. In his book *Economics: The User's Guide*, Chang argues that many economists believe, and tell other people, that economics is a 'value-free' science, like physics or chemistry. However, Chang argues that economics is a fundamentally political and moral subject in relation to choosing the best option. Whereas the particles and compounds studied by natural scientists do not hold political and moral views, human beings who populate the economy do, and so we cannot fully understand the economy without understanding politics and ethics. No economic argument can be free from politics and ethics: for example, the economic case for trying to persuade people to buy fair-trade products.

Very often economists respond to the criticism that their subject is 'soft' by arguing that they are only concerned with 'positive economics', which they claim is based on quite strict use of scientific methodology. Positive economics is concerned with 'what is' and 'what will happen' if a course of action is taken or not taken. In contrast, 'normative economics' is concerned with 'what *should* or *ought* to be'.

TEST YOURSELF 1.3
The study of economics can be helped by looking at other relevant social sciences. Can you name three?

The difference between positive and normative statements

KEY TERMS

normative statement a statement that includes a value judgement and cannot be refuted just by looking at the evidence.

positive statement a statement of fact that can be scientifically tested to see if it is correct or incorrect.

A lot of economics is concerned with what people *ought* to do. This is particularly true of the government. Should the government try to reduce unemployment, control inflation and achieve a 'fair' distribution of income and wealth? Most people probably think that all these objectives are desirable. However, they all fall within the remit of normative economics. Normative economics is about value judgements and views, but because people have different views about what is right and wrong, **normative statements** cannot be scientifically tested.

By contrast, a **positive statement** can be tested to see if it is incorrect. If a positive statement does not pass the test, it is falsified. A positive statement does not have to be true, however. For example, the statement that the Earth is flat is a positive statement. Although once believed to be true, the statement was falsified with the growth of scientific evidence. The key point is that positive statements can in principle be tested and possibly falsified, while normative statements cannot. Normative statements include ethical, or moral, judgements. Words such as *ought*, *should*, *better* and *worse* often provide clues that a statement is normative.

To take an economic example, consider the statement 'If the state pension were to be abolished, a million older people would die of hypothermia.' This is a positive statement which could be tested, though few if any people would want to do this. By contrast, the statement 'The state pension ought to be abolished because it is a waste of scarce resources' is normative, containing an implicit value judgement about the meaning of the word 'waste'.

STUDY TIP

Make sure you understand fully the difference between a positive and a normative statement. These concepts are often tested as multiple-choice questions in the exam and you will need to be able to select which options are most clearly positive or normative statements. Remember that normative statements often include clues that they are opinions, by using words or phrases such as 'should', 'ought to' or 'must'.

TEST YOURSELF 1.4

Which one of the following is an example of a normative economic statement?

A Increased use of diesel-engine cars will lead to more atmospheric pollution.

B Higher taxes always lead to higher prices.

C The government should spend more on roads.

D Consumers generally act rationally.

How value judgements influence economic decision making and policy

Economists emphasise the distinction between normative and positive economics, but they often forget that the decision to study one over the other is itself a value judgement, and therefore a normative decision. A value judgement

is about whether something is desirable or not — if we believe it is more desirable to study what *is* happening in the economy rather than what *ought* to happen, we have made a value judgement. Economics necessarily requires that government ministers make value-based judgements when deciding on economic policies. Despite this, economists often wrongly insist that the subject is value-judgement free.

One particular chancellor of the exchequer (the UK government minister in overall charge of economic policy) once said: 'Rising unemployment and the recession have been the price that we have had to pay to get inflation down. That price is well worth paying.' Government ministers are often less frank than this, knowing that their political opponents and the media will immediately seize on the argument that those in power are uncaring and cynical people. However, the quote does serve to illustrate how decision-makers make value judgements when making economic policy decisions.

On occasion, government ministers make decisions on issues such as where a new airport should be located or whether high-speed trains are worthwhile. Before making decisions on issues like these, the policy-makers know in advance that large groups of the population will strongly oppose whatever decision is eventually made. To try and avoid public hostility, government ministers usually create the illusion that the decision-making process is completely scientific and objective. To do this, they hire independent 'experts' to provide advice. But the choice of expert in itself involves a value judgement. Do you choose someone you know in advance is sympathetic to the government's cause, or are you more willing to go for someone more impartial? Whichever way you go, the so-called scientific processes used by the 'experts' to reach their conclusions may be full of value judgements. A classic case involved weighing up the costs and benefits of the location of an additional London airport, which ultimately depended on putting money values on an hour of a business person's time, and an hour of a holidaymaker's time. It was quickly found that when different values were put on these, the airport location recommended by the experts would have 'lost out' under different costing criteria.

STUDY TIP

Make sure you understand fully how value judgements link to normative statements. Value judgements concern the desirability, or otherwise, of things. The fields of politics and economic policy making are based on value judgements. See if you can identify the normative statements made by politicians on news programmes.



Government priorities and spending change — each decision has consequences

The impact of moral and political judgements

Whatever decision is eventually made in the course of framing government economic policy, there will always be winners who gain and losers who suffer as a result of the decision. Governments often claim they have a moral right to make such decisions. They argue that their political manifesto published before the previous general election gives them the mandate, supported by the voters, to carry out their policies, regardless of the fact that among the electorate there will inevitably be some losers.

One example is provided by former US president Donald Trump's decision in 2018 to implement a 'zero tolerance' policy on illegal immigration into the USA. The policy involved splitting adult illegal migrants from their children and keeping them in wire cages along the US border with Mexico. On humanitarian grounds the policy was so unpopular in the USA that it was quickly abandoned in a policy 'U-turn', though the other parts of Trump's anti-immigrant policy remained in place.

SECTION 1.1 SUMMARY

- Economic methodology involves the application of tested economic theories to explain real-world economic behaviour.
- It is important to understand the difference between positive and normative statements. A positive statement can be tested to see if it is correct or false; a normative statement is a statement that includes a value judgement which cannot be refuted purely by looking at the evidence.
- A value judgement is about whether something is desirable or not.
- Government ministers make value-based judgements when deciding on economic policies.
- There will always be winners who gain and losers who suffer as a result of government policy decisions.

1.2 The nature and purpose of economic activity

Needs and wants

KEY TERMS

need something that is necessary for human survival, such as food, clothing, warmth or shelter.

want something that is desirable, such as a mobile phone, but is not necessary for human survival.

The production of goods and services to satisfy people's needs and wants is the central purpose of economic activity. A **need** is something people must have, something that they cannot do without. Food provides an example. If people starve, they will eventually die. By contrast, a **want** is something people would *like* to have, but which is not essential for survival. It is not absolutely necessary, but it is a good thing to have. Books provide an example. Some people might argue that books are a need because they think they can't do without them. But they don't need literature to survive. They do need to eat. (It is worth noting that food can be both a need and a want, depending on the type of food. Protein and vitamins are needs, but bars of chocolate are wants. People don't need to eat chocolate to survive.)

KEY TERM
economic welfare the economic well-being of an individual, a group within society, or an economy.

Satisfying people's needs and wants means improving **economic welfare**. Welfare is a concept discussed a lot by economists, but often without a clear indication of what the concept means. Welfare is anything that makes a person happier and improves their economic well-being. Short-term happiness may be at the expense of long-term well-being. The consumption of more material goods and services *usually* improves economic welfare, though, in the long term, consuming more and more food, and the wrong type of food, can lead to health problems.

STUDY TIP
Make sure you can explain the meaning of the term 'economic welfare'. This concept is important throughout your studies because, for example, it is the benchmark by which economists judge whether a policy is worth implementing.

There are also important elements of human happiness and welfare that have nothing to do with the consumption of material goods. These include quality-of-life factors, such as the pleasure gained from family and friends or from contemplating a beautiful view.

TEST YOURSELF 1.5
With one example of each, explain the difference between a need and a want.

The key decisions of what and how to produce

KEY TERM
economic system the set of institutions within which a community decides what, how and for whom to produce.

We mentioned at the beginning of this chapter that decisions are made about what to produce, how to produce and who is to benefit from the goods and services produced. How these decisions are made depends upon the nature of the **economic system** within the economy. The set of institutions within which a community decides what, how and for whom to produce is called an economic system. Although the problem of scarcity is fundamental and common to all forms of human society, from the indigenous peoples in the Amazonian rainforest, to wealthy nations such as the United States of America, different economic systems have evolved in different societies.

SYNOPTIC LINK
The key economic concept of scarcity is explained in section 1.4 of this chapter, and also in section 1.3, in the context of scarce environmental resources. The existence of scarcity means that societies must make sometimes difficult choices about what products are made, how they are made and who will receive them.

Perhaps the most widely used method of defining and classifying economic systems is according to the way or mechanism through which scarce resources reach the people who eventually consume or use them. Although there are varied ways in which wealth and purchasing power can be allocated among individuals, including inheritance and other types of gift, theft and

KEY TERMS

market economy an economy in which goods and services are purchased through the price mechanism in a system of markets.

command economy (also known as a **planned economy**) an economy in which government officials or planners allocate economic resources to firms and other productive enterprises.

mixed economy an economy that contains both a large market sector and a large non-market sector in which the planning mechanism operates.

TEST YOURSELF 1.6

To what extent is the UK economy a mixed economy?

luck or chance, such as winning a fortune on the National Lottery, the two mechanisms by which economic systems are defined are the market mechanism (or price mechanism) and the command mechanism (or planning mechanism). An economic system in which goods and services are purchased through the price mechanism in a system of markets is called a **market economy**, whereas one in which government officials or planners allocate economic resources to firms and other productive enterprises is called a **command economy** (or planned economy).

Between these extremes, many economies, particularly those in developed economies such as the UK, are called **mixed economies**. A mixed economy, as the name suggests, is a mixture of different types of economic system. A mixed economy contains both a large market sector and a large non-market sector in which the planning mechanism operates. Figure 1.2 illustrates mixed economies in relation to planned and market economies.

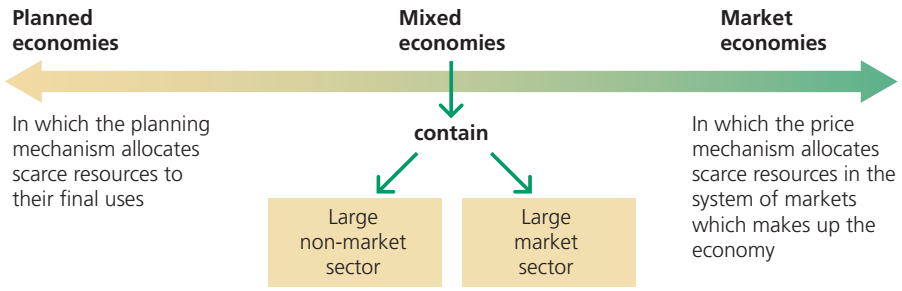


Figure 1.2 Command economies, mixed economies and market economies

APPLICATION OF ECONOMICS IN THE REAL WORLD 1.1

Allocating resources through the price mechanism and the planning mechanism

In a pure market economy, the price mechanism performs the central economic task of allocating scarce resources among competing uses through the markets which make up the economy. Transport costs and lack of information may create barriers that separate or break up markets. In past centuries, such barriers often prevented markets from operating outside the relatively small geographical area of a single country or even a small region within a country.

However, while some markets exist in a particular geographical location — for example, a street market or until quite recently the London Stock Exchange — many markets do not. In recent years, modern developments have allowed goods to be transported more easily and at lower cost, and have helped in the transmission of market information via telephone and the internet. This has enabled many markets, especially commodity and raw material markets and markets in financial services, to become truly global or international markets functioning on a worldwide basis.

A complete command economy is an economy in which all decisions about what, how, how much, when, where and for whom to produce are taken by a central planning authority, issuing commands or directives to all the households and producers in the society. Such a system could only exist within a very rigid and controlled political framework because of the restrictions on individual decision making that are implied.

In much the same way as a pure market economy, in which the price mechanism alone allocates resources, is a theoretical abstraction, so no economy in the real world can properly be described as a complete or pure planned economy. Before the collapse of the communist political system around 1990, some countries in eastern Europe were centrally planned economies. However, they were not pure planned economies. Production but not consumption was planned. Consumers often had to queue to get consumer goods, whose prices were fixed by the planners. Shortages resulted, which, together with the generally inferior quality of consumer goods, contributed to the breakdown of the command economies.

Some communist countries still exist, namely the People's Republic of China, North Korea, Vietnam and Cuba. However, all these countries, with the exception until recently of North Korea, have encouraged the growth of markets to a greater or lesser extent. They have communist political systems, but they have moved away from being pure command economies. In a sense, China is now more capitalist than the USA and the mixed economies of western Europe. The term 'state capitalism' is now used to describe much of China's economy.

Follow-up questions

1 What are the differences between a pure market economy and the market sector of a mixed economy?
2 What is the other name of a planned economy?
3 With the help of Chapter 3, explain how the price mechanism allocates scarce resources between competing uses.
4 Why did the command economies of central and eastern Europe, such as the Union of Soviet Socialist Republics (USSR) break up?

APPLICATION OF ECONOMICS IN THE REAL WORLD 1.2

The UK as a mixed economy

The UK economy developed into a mixed economy after the Second World War ended in 1945, when a number of important industries such as coal, rail and steel were nationalised and taken into public ownership. Previously, the 1944 Education Act had extended state provision of education, and the creation of the National Health Service in 1948 did the same for healthcare.

For about 30 years after the end of the Second World War, from the 1940s to the 1970s, the majority of UK citizens (and the major political parties) agreed that the mixed economy was working well. Most people believed that certain types of economic activity, particularly the production and distribution of consumer goods and services, were best suited to private enterprise and the market economy. But people also accepted that utility industries such as gas and electricity should be nationalised, and that important services such as education, healthcare and roads should in part be provided by government, outside the market, and financed through the tax system. In short, a consensus existed around the belief that the mixed economy was right for the UK.

However, from about 1980 onwards, many economists and politicians began to blame the mixed economy for the UK's deteriorating economic performance relative to that of its main competitors in western Europe and Japan. Critics argued that the public and non-market sectors of the economy were inefficient and wealth consuming rather than wealth creating. They felt that these sectors had become too big and that their dominance should be weakened. Critics of the mixed economy argued that a concerted effort should be made to change the nature of the UK economy fundamentally, by increasing private ownership and market production.

Successive governments implemented policies that changed the nature of the mix in favour of private ownership and market forces, at the expense of public ownership and state planning. The UK economy is now much closer to being a pure market and private enterprise economy than it was 45 years ago. The three main policies used to change the nature of the UK economy have been privatisation, marketisation and deregulation, policies which collectively can be called 'economic liberalisation'.

Privatisation involved the sale of state-owned assets such as nationalised industries to private owners. This was often accompanied by marketisation (or commercialisation), whereby prices are charged for goods and services that the state previously provided free of charge. Deregulation, the third aspect of liberalisation, attempts to remove barriers to entry (see Chapter 5) and government red tape and bureaucracy from the operation of markets.

Follow-up questions

1 Critics of the UK economy have called it a 'mixed-up' economy rather than a 'mixed economy'. What do you think they mean?
2 What is meant by the term 'private enterprise'?
3 What is the difference between privatisation and marketisation?
4 Do you think that there is now a case for increasing rather than reducing state ownership of UK industry? Justify your answer.

SECTION 1.2 SUMMARY

- A need is something people must have; it is something that they cannot live without.
- A want is something people would like to have, but which is not essential for survival.
- How decisions are made about what to produce, how to produce and who is to benefit from the goods and services produced depends upon the nature of the economic system.
- Market economies, command economies and mixed economies are examples of economic systems, defined in terms of whether or not the price mechanism or the planning mechanism allocates scarce resources among competing uses.
- Economic systems can also be defined in terms of who owns the means of production.
- The UK economy is often called a mixed economy, being on the one hand a mix of market and non-market sectors and on the other hand a mix of privately owned and publicly owned sectors.

1.3 Economic resources

KEY TERMS

production converts inputs or factor services into outputs of goods and services.

capital good (also known as a **producer good**) a good which is used in the production of other goods or services.

consumer good a good which is consumed by individuals or households to satisfy their needs or wants.

For most people, most of the time, increased consumption of material goods is an important part of improving economic welfare. Most of the goods we consume must first be produced. This requires the use of economic resources. These goods are scarce in relation to demand, which gives rise to the need for rationing and economising in their use.

The basic nature of **production** is shown in Figure 1.3. Production is a process, or set of processes, that converts inputs into outputs. The eventual outputs are the consumer goods and services that go to make up our standard of living, though inputs are of course also used to produce the **capital goods** that are necessary for the eventual production of **consumer goods**.

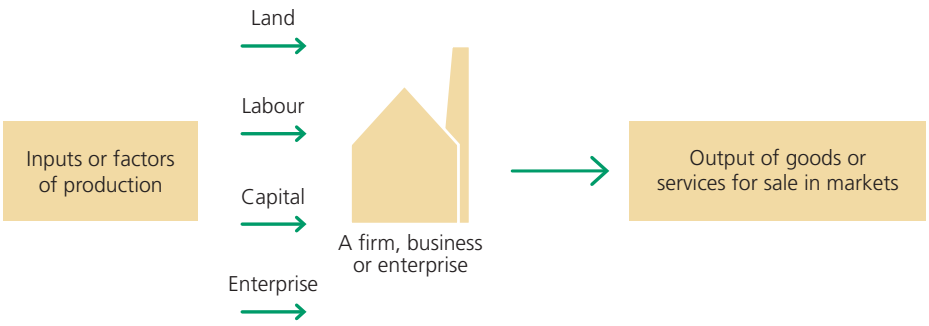


Figure 1.3 The basic nature of production

The factors of production

KEY TERM

factors of production inputs into the production process (e.g. land, labour, capital and enterprise).

Economists call the inputs into the production process, which are shown in Figure 1.3, the **factors of production**. Four factors of production are usually identified. These are land, labour, capital and enterprise, the last often being called the entrepreneurial input.

Entrepreneurs are different from the other factors of production. They are the people who address the issues introduced earlier, deciding what to produce,

how to produce it and for whom to produce it. An entrepreneur decides how much of the other factors of production, including labour, to employ. The costs of employing land, labour and capital, together with the cost of the entrepreneur's own services, become the firm's costs of production. In essence, the entrepreneur is a financial risk-taker and decision-maker. Profit, which is the entrepreneur's financial reward, results from successful decision making. Entrepreneurial profit is the profit left over after the cost of employing the other factors of production is deducted from the sales revenue gained from selling the goods and services the entrepreneur decides to produce.

STUDY TIP

It is important to understand that factors of production are *inputs* used to produce *output* of goods and services. The factor of production known as enterprise is usually seen as the factor which coordinates factor inputs of capital, land and labour in pursuit of profit.

The environment as a scarce resource

KEY TERMS

renewable resource a resource (e.g. timber) that with careful management can be renewed as it is used.

non-renewable resource (also known as a **finite resource**) a resource (e.g. oil) that is scarce and runs out as it is used.

Environmental resources comprise all the natural resources that are used or can be used in the economic system. These are:

- physical resources, such as soil, water, forests, fisheries and minerals
- gases, such as hydrogen and oxygen
- abstract resources, such as solar energy, wind energy, the beauty of the landscape, clean air and water.

Environmental resources can be split into renewable and non-renewable resources, with the latter further divided into recyclable and non-recyclable resources. **Renewable resources**, which are reproducible and perpetually maintainable, include forests, animals and water. The availability of these resources depends, however, on their management by humans. By contrast, **non-renewable resources**, such as oil, gas and minerals, cannot be regenerated or their regeneration is so slow that the stock of resources cannot meaningfully be increased. These are **finite resources**. Recyclable non-renewable resources such as minerals, paper and glass can be reused in the economic system. In theory, all of these resources can be recycled but it is not always possible and economic to recycle more than a small fraction. Non-recyclable resources such as coal, gas and oil are finite in the sense that, once used, their stock is no longer available for future use.

Environmental resources are part of the factor of production, land. Some environmental resources, such as the air we breathe and the water we drink, are often described as the 'free gifts of nature'. However, this view can be questioned. In most countries and regions where large numbers of people live, clean air and drinkable water are scarce commodities and not the 'free gifts of nature'. The need to get rid of the effects of pollution created by humankind means that clean air and water are scarce and not free. Resources which could be put to other uses are used instead to produce clean air and water. Production and consumption activities taking place in the economy affect and often damage the natural environment.



Oil is a non-renewable resource

TEST YOURSELF 1.7

Distinguish between a non-renewable and a renewable environmental resource.

SECTION 1.3 SUMMARY

- Increased consumption of material goods is an important part of improving economic welfare.
- Most of the goods we consume must first be produced.
- This requires the use of economic resources.
- These goods are scarce in relation to demand, which gives rise to the need for rationing and economising in their use.
- Production is a process, or set of processes, that converts inputs into outputs.
- Inputs are called factors of production and are economic resources.
- Land, labour, capital and enterprise (often called the entrepreneurial input) are the factors of production.
- The entrepreneur is the factor of production that decides what to produce, how to produce and for whom to produce.
- Inputs are also used to produce the capital goods that are necessary for the eventual production of consumer goods and services.
- Environmental resources comprise all the natural resources that are used or can be used in the economic system.

1.4 Scarcity, choice and the allocation of resources

Scarcity: the fundamental economic problem

We have already mentioned that we live on a finite planet in which most economic resources are limited. As we have seen, their use usually has to be rationed either by the price mechanism or by the planning mechanism.

The **fundamental economic problem** is therefore **scarcity**. In a world of scarcity, people (even the very rich) have limited incomes, which means that they face a budget constraint. When the price mechanism is rationing scarce resources between competing uses, a budget constraint represents all the combinations of goods and services that a consumer may purchase given current prices and their limited income.

KEY TERMS

fundamental economic problem how best to make decisions about the allocation of scarce resources among competing uses, so as to improve and maximise human happiness and welfare.

scarcity results from the fact that people have unlimited wants but resources to meet these wants are limited. In essence, people would like to consume more goods and services than the economy is able to produce with its limited resources.

The need for choice

KEY TERM

choice choosing between alternatives when making a decision on how to use scarce resources.

If goods are scarce and incomes are limited, a **choice** or choices have to be made. Consider, for example, a family with a weekly income of £1,200. The family currently spends £350 on housing, £350 on food, £300 on other goods and services, including heating and lighting, and £100 on entertainment. The family's total weekly spending on goods and services is thus £1,100, meaning the family manages to save £100. Suddenly, the cost of housing rises to £550. To avoid getting into debt, and assuming that family income cannot increase, one or more probably unpleasant choices will have to be made. An obvious possibility is to cut down on entertainment, such as visits to the cinema. Other possibilities could be spending less on home heating, buying cheaper food, reducing the purchase of non-essential items and stopping saving. Something will have to be given up. Unless the family gets into debt or its income increases, it will have to economise on its spending and saving decisions.

It is important to appreciate that, even without an increase in house prices, scarcity means that individuals and households are constantly making choices about how to spend their limited incomes and how to make the best use of their time. A decision to spend more on a holiday, for example, means that a family chooses to spend less on other goods, or to save less.

Choices have an opportunity cost

KEY TERM

opportunity cost the cost of giving up the next best alternative.

A need for choice arises whenever an economic agent (e.g. an individual, a household or a firm) has to choose between two or more alternatives which are mutually exclusive, in the sense that it is impossible or impractical to achieve both at the same time. In the jargon of economics, an **opportunity cost** is involved.

If you ask friends who haven't studied economics the meaning of the word 'cost', typically they will answer that cost is the money cost either of producing a good or of buying a good. Economists, by contrast, focus on opportunity cost. The opportunity cost of any choice, decision or course of action is measured in terms of the alternatives that have to be given up.

TEST YOURSELF 1.8

What is meant by rational economic behaviour?

Economists generally assume that people behave rationally. Rational behaviour means people try to make decisions in their self-interest or to maximise their

SYNOPTIC LINK
The assumption that people are always rational when making choices is questioned in the discussion of behavioural economics in Chapter 2.

private benefit. When a choice has to be made, people always choose what they think at the time is the best alternative, which means that the second best or next best alternative is rejected. For rational people, the opportunity cost of any decision or choice is the next best alternative sacrificed or forgone. For example, if you choose to spend half an hour watching a TV programme, the opportunity cost is the lost opportunity to spend this time reading a magazine or book.

TEST YOURSELF 1.9
Can you explain one factor that in real life may prevent a person from making a rational decision?

STUDY TIP
Make sure you can link together the three concepts of scarcity, choice and opportunity cost. As economic resources are finite, society must make choices about how to allocate them. In doing so, there will always be alternatives which are not chosen that would have given some benefit to society; this is the concept of opportunity cost.

Rational behaviour and opportunity cost

Look carefully at the sentence in the previous paragraph about opportunity cost: *When a choice has to be made, people always choose what they think at the time is the best alternative, which means that the second best or next best alternative is rejected.* The words 'think at the time' are quite important. Suppose you decide to see a film, believing at the time you are going to enjoy the film. Two hours later, coming out of the cinema, you say 'That was a load of rubbish, I wish I hadn't bothered.' Nevertheless, at the time the decision was made, deciding to watch the film was a rational decision because you believed you would enjoy watching it. Having left the cinema, irrational behaviour would be going back to the box office and paying good money to see the film again.

Even if the film can be watched freely on television, there is still an opportunity cost, though in this case it only involves time. Time is scarce, and the two hours spent watching the film cannot be spent on some other activity such as reading a book. A choice has to be made between the two activities. Of course, the way the choice is exercised will differ for different people, even if everyone enjoys both watching films and reading. People with plenty of time to spare — for example, retired or unemployed people — value time less highly than a financial trader working in a dealing room from 7 a.m. until 9 p.m.

TEST YOURSELF 1.10
Give an example of an opportunity cost facing an A-level student when choosing to spend a whole evening playing computer games.

Going to a cinema to watch a film is, of course, a consumer activity. People have to think carefully about how they spend their limited incomes. In economics this is known as a budget constraint. Money spent on one good cannot be spent on another good. But there is always an opportunity cost when deciding to consume a good, involving time if not money.

Firms also have to make choices about what and how to produce. Consider a textile manufacturer which can produce either shirts or dresses from the same production line, but not both goods at the same time. In this situation, the opportunity cost of producing more shirts is the number of dresses sacrificed or forgone. Suppose also that both shirts and dresses can be produced using one of two different technologies. These are a labour-intensive technology involving lots of workers but very little capital equipment, and a capital-intensive technology in which there are very few workers but expensive

automated capital equipment. Given the budget constraint facing the firm, the opportunity cost of choosing one method of production is the sacrificed opportunity to use the other method.

A further example of opportunity cost arises when a teenager makes a decision about whether to leave school and get a job, or to go to university. Very often this involves the choice between income now and income in the future. Economists call this inter-temporal choice, or choice over time.

QUANTITATIVE SKILLS 1.1

Worked example: calculating an opportunity cost

A small electrical goods manufacturer can produce either TV sets or radio sets using all its available resources. Table 1.1 shows the different combinations of the two goods the firm can produce.

Table 1.1 Production possibilities for TVs and radio sets

TV sets	Radio sets
0	30
1	29
2	27
3	24
4	20
5	15
6	9
7	0

What happens to the opportunity cost of TV sets in terms of radios, as TV set production increases from 0 to 7 sets?

If the firm chooses to produce only 1 TV set, its opportunity cost is 1 radio set forgone (30 – 29 radio sets). Performing a similar calculation when TV set production is increased by an extra unit, the opportunity cost of the second TV set is 2 radio sets. All the opportunity costs are set out in Table 1.2.

Table 1.2 Opportunity costs of producing an extra TV set

1st TV set	1 radio set (30 - 29)
2nd TV set	2 radio sets (29 - 27)
3rd TV set	3 radio sets (27 - 24)
4th TV set	4 radio sets (24 - 20)
5th TV set	5 radio sets (20 - 15)
6th TV set	6 radio sets (15 - 9)
7th TV set	9 radio sets (9 - 0)

The data show an increasing opportunity cost in terms of radio sets forgone as production of TV sets increases.

STUDY TIP

Make sure you understand and can distinguish between the words 'micro' and 'macro'. 'Micro' means small, whereas 'macro' means large. Microeconomics, which is covered in the first eight chapters of this book, looks at the 'little bits' of the economy: for example, individual markets, firms and consumers. By contrast, macroeconomics, covered in Chapters 9-15, looks at the 'aggregate economy' or the economy 'as a whole'.

SECTION 1.4 SUMMARY

- The economic problem is how limited resources are used in relation to people's desires and wants.
- The economic problem results from relative scarcity.
- Scarcity results in the need for choice.
- Whenever a choice has to be made, there is an opportunity cost.
- The opportunity cost of any decision is the next best alternative forgone.
- Economists generally assume that people are rational, choosing the best alternative available.

1.5 Production possibility diagrams

KEY TERMS

production possibility frontier a curve depicting the various combinations of two products (or types of product) that can be produced when all the available resources are fully and efficiently employed.

technical progress new and better ways of making goods and new techniques for producing more output from scarce resources.

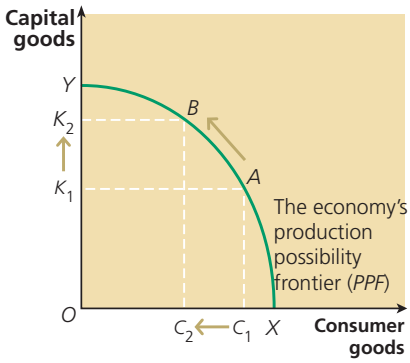


Figure 1.4 Using a production possibility frontier diagram to illustrate the economic problem

So far, we have focused mainly on how scarcity and choice may affect firms, families and individuals at the microeconomic level. In much the same way, but on a far grander scale, the economy of the nation as a whole faces a similar need for choice. To explain how the economic problem affects the whole economy, we will use a diagram which you will come across again and again in your economics course — a production possibility diagram.

The key feature of a production possibility diagram is a **production possibility frontier (*PPF*)** or production possibility curve. A *PPF* illustrates the different combinations of two goods, or two sets of goods, that can be produced with a fixed quantity of resource, providing we assume that all available resources are being utilised to the full. The *PPF* in Figure 1.4 illustrates the different combinations of capital goods and consumer goods that the whole economy can produce when all the economy's resources are employed, with no spare capacity. To put it another way, the *PPF* shows what the economy can produce, assuming that all the labour, capital and land at the country's disposal are employed to the full, and assuming a given state of **technical progress**.

Given that resources and capacity are limited, a choice has to be made about the type of good to produce. Look closely at points *X* and *Y* on the diagram. Point *X* shows the maximum possible output of consumer goods, assuming that the economy only produces consumer goods (i.e. no capital goods are produced). Likewise, point *Y* shows the maximum possible output of capital goods, assuming that the economy only produces capital goods. In fact, points *X* and *Y* show the two extreme production possibilities, since all goods are either consumer goods or capital goods. Finally, the line drawn between points *X* and *Y* in Figure 1.4 is the economy's production possibility frontier. The *PPF* shows all the different combinations of consumer goods and capital goods that can be produced, given the assumptions mentioned earlier about full employment of available resources and the state of technical progress. Point *A*, for example, shows K_1 capital goods and C_1 consumer goods being produced. An increase in capital goods production to K_2, shown at point *B*, means that consumer goods production falls to C_2. $C_1 - C_2$ is the opportunity cost of producing $K_2 - K_1$ additional capital goods. Whichever combination of capital and consumer goods is actually chosen reflects decisions made in society about allocating scarce resources between competing uses.

STUDY TIP

You must learn to draw and interpret production possibility diagrams, which are important in both microeconomics and macroeconomics. At the micro level they can be used to illustrate scarcity, choice, opportunity cost and productive efficiency. At the macro level, they can be used to illustrate economic growth, full employment and unemployment.

QUANTITATIVE SKILLS 1.2

Worked example: calculating opportunity cost: apples and pears

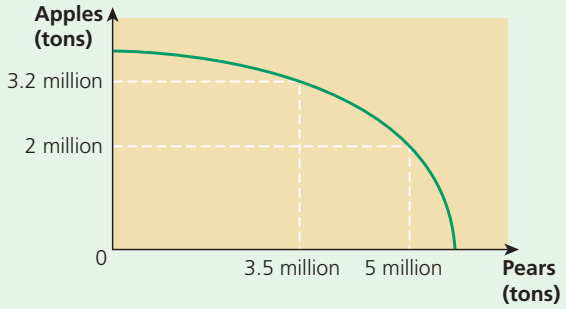


Figure 1.5 Production possibilities for apples and pears

A country is currently producing 3.2 million tons of apples and 3.5 million tons of pears. Calculate the opportunity cost in terms of apples of increasing pear production to 5 million tons.

The opportunity cost of increasing pear production by 1.5 million tons is the sacrifice of 1.2 million tons of apples.

SYNOPTIC LINK

As we explain below, production possibility diagrams can also be used to illustrate economic growth. This is revisited in Chapter 11. Chapter 10 explains another way of illustrating economic growth, in terms of aggregate demand and aggregate supply or *AD/AS* analysis. *AD/AS* analysis is one of the key models in macroeconomics, which is concerned with the performance of the entire economy. In contrast, microeconomics is concerned with individual markets within an economy.

Macroeconomic *PPF* diagrams

The production possibility frontier in Figure 1.4 shows the economy *as a whole*, which means that the *PPF* depicts the macroeconomy. Before we look at some microeconomic production possibility frontiers, we will explain two important ways in which macroeconomic *PPF* diagrams can be used.

Using a *PPF* diagram to show economic growth

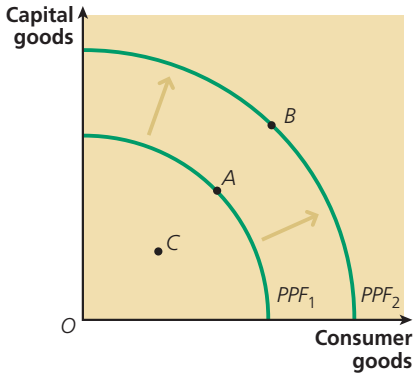


Figure 1.6 Long-run and short-run economic growth

Figure 1.6 shows how a *PPF* diagram can be used to illustrate **economic growth**. There are two forms of economic growth, which are explained in detail in Chapter 11. These are long-run economic growth and short-run economic growth.

Economic growth is defined as the increase in the *potential* level of real output the economy can produce over a period of time: for example, a year. Strictly, this is long-run economic growth, which is not the same as short-run economic growth. If the economy's production possibility frontier is PPF_1 initially, short-run economic growth is shown by the movement from point *C inside* the frontier to a point, such as point *A*, *on* the frontier.

Long-run economic growth is shown by the *outward* movement of the frontier to PPF_2. The movement from point *A* to point *B* depicts long-run economic growth. Short-run growth makes use of spare capacity and takes up the slack in the economy, whereas long-run growth increases total productive capacity.

KEY TERM

economic growth the increase in the *potential* level of real output the economy can produce over a period of time.

Using a *PPF* diagram to show full employment and unemployment

As mentioned earlier, all points on a production possibility frontier show full employment of available resources. For a macroeconomic production

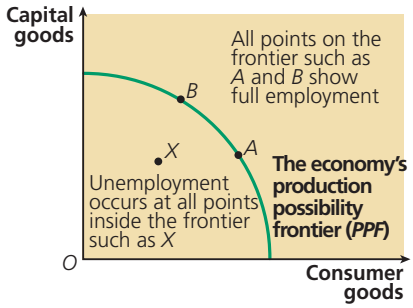


Figure 1.7 Using a production possibility curve diagram to show full employment and unemployment in the economy

possibility frontier, this means **full employment** of labour as well as other resources that can be used in the course of production. Thus, points *A* and *B* in Figure 1.7 show full employment when the economy's production possibility frontier is in the position indicated.

By contrast, a point *inside* the *PPF*, such as point *X*, shows that some resources, including labour, are not being employed. There is unused capacity in the economy. (Note that if long-run economic growth were to move the *PPF* outwards to a new position, points *A* and *B* on the 'old' production possibility frontier would now be points of **unemployment**, as they would be located inside the new 'further out' frontier.)

KEY TERMS

full employment when all who are able and willing to work are employed.

unemployment when not all of those who are able and willing to work are employed.

TEST YOURSELF 1.11

Illustrate full employment and unemployment on a production possibility frontier diagram.

SYNOPTIC LINK

PPF diagrams can be used to illustrate a range of macroeconomic concepts such as full employment and unemployment. These concepts feature in Chapters 11, 12 and 13, where they are also illustrated with aggregate demand and aggregate supply (*AD/AS*) diagrams.

Microeconomic *PPF* diagrams

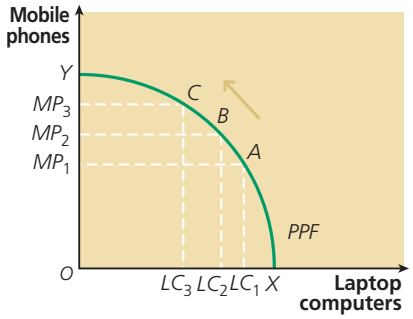


Figure 1.8 The production possibility frontier facing a firm producing mobile phones and laptop computers

Look carefully at the microeconomic production possibility frontier in Figure 1.8. We can tell this is a *microeconomic* diagram because of the labels on the two axes of the graph. The diagram depicts a situation in which a firm can produce both mobile phones and laptop computers from the resources it has available. As in Figure 1.4, points *X* and *Y* show the extreme possibilities facing the firm. Point *X* on the horizontal axis shows the maximum possible output of laptop computers when no mobile phones are produced. This means that all available resources are devoted to the production of laptop computers.

Conversely, point *Y* on the vertical axis shows the opposite situation: the maximum possible output of mobile phones when zero laptop computers are produced. As was the case in the macroeconomic production possibility diagrams, all points on the *PPF* between *X* and *Y* show different combinations of the two goods being produced between the two extreme possibilities.

We can use Figure 1.8 to explain a number of important economic relationships.

KEY TERM

resource allocation the process through which the available factors of production are assigned to produce different goods and services (e.g. how many of society's economic resources are devoted to supplying different products such as food, cars, healthcare and defence).

Scarcity, resource allocation and choice (revisited)

A *PPF* diagram such as Figure 1.8 shows different possible ways in which scarce resources can be allocated between competing uses. This involves choice. Compared to the **resource allocation** at point *A*, point *B* shows the effect of shifting more resources into the production of mobile phones, with fewer resources being allocated to laptop computer production.

Opportunity cost (revisited)

Production possibility diagrams provide a very good way of illustrating opportunity cost. Suppose, for example, that the firm in Figure 1.8 initially produces MP_1 mobile phones and LC_1 laptop computers. This combination of the two goods is shown at point *A* in the diagram. A decision by the manufacturer to increase production of mobile phones from MP_1 to MP_2 means that computer production falls by $LC_1 - LC_2$. Moving from point *A* to point *B* on the curve, the fall in computer production is the opportunity cost of the increase in phone production.

Look now at the shape of the *PPF* in Figure 1.8. You can see that the slope of the curve falls, moving up the curve from point *A* to point *B*, and indeed to all other points further up the curve. There is a reason for this. The slope shows the opportunity cost of producing more mobile phones in terms of the laptop computers that have to be sacrificed. When mobile phone production increases from MP_1 to MP_2 laptop computer production falls by $LC_1 - LC_2$. This is the opportunity cost involved. But suppose mobile phone production increases again by the same amount as before (which means that $MP_3 - MP_2$ is the same as $MP_2 - MP_1$). In this situation, shown at point *C* on the curve, more laptop computers than before have to be given up. $LC_3 - LC_2$ is larger than $LC_2 - LC_1$. The slope of the curve of the production possibility frontier shows that the opportunity cost of producing mobile phones increases as more mobile phones are produced. A greater number of laptop computers have to be sacrificed whenever an extra mobile phone is produced.



A large company such as Apple must decide how to allocate its resources

TEST YOURSELF 1.12
Provide another example of a production possibility frontier diagram illustrating a microeconomic idea or concept.

SYNOPTIC LINK
Link the microeconomic use of production possibility frontier diagrams to the coverage of the price mechanism in Chapter 3 and to production, costs and revenue in Chapter 4. All of these models can be used to illustrate the concepts of efficiency and choice.

Productive efficiency and production possibility diagrams

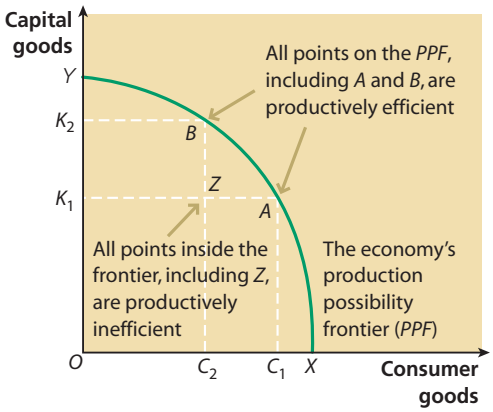


Figure 1.9 Productive efficiency and the economy's *PPF*

You will come across the concept of economic efficiency on numerous occasions as you proceed through the course and this book. There are a number of different measures of economic efficiency, one of which is **productive efficiency**. Chapter 5 explains productive efficiency in terms of a firm minimising the average cost of producing a good. In this chapter, we focus on two other ways of explaining the concept, both of which are illustrated by the production possibility curve in Figure 1.9.

Productive efficiency occurs when output is maximised from available inputs. But we know already that a *PPF* shows maximisation of output from available inputs at every point on the curve, though the combination of the two goods (in this case, capital goods and consumer goods) varies at different points on the curve. This means that all points *on* the economy's production possibility frontier shown in Figure 1.9, including points *A* and *B*, are productively efficient. By contrast, all points *inside* the *PPF* are productively *in*efficient, including point *Z*. Productive inefficiency occurs when output is not maximised from available inputs. At point *Z*, the economy is not employing all the available resources, including, of course, labour. Productive inefficiency is often associated with unemployment.

Consider now a movement *along* the *PPF* from point *A* to point *B*. The diagram is telling us that, when on the *PPF*, more capital goods can be produced only by giving up the production of some consumer goods. This is another way of explaining productive efficiency. Productive efficiency occurs when producing more of one good involves reducing production of other goods. By contrast, when the economy is productively inefficient at point *Z*, more capital goods *and* consumer goods can be produced by taking up the slack in the economy and making use of idle resources.

KEY TERM

productive efficiency for the economy as a whole occurs when it is impossible to produce more of one good without producing less of another. For a firm it occurs when the average total cost of production is minimised.

TEST YOURSELF 1.13

A production possibility frontier can be used to show all the following **except**:

A all the different combinations of two goods that can be produced from available inputs

B all the productively efficient combinations of output that can be produced from available inputs

C the opportunity cost of producing more of a particular good

D the economic welfare a consumer gains from different combinations of goods

SYNOPTIC LINK

As noted, there are various types of economic efficiency you need to learn and understand as you proceed through the course. One type is productive efficiency, first introduced in this chapter, which can be illustrated on a production possibility frontier diagram. Other types of efficiency are best learnt later in the course. One of these is allocative efficiency. Chapter 5 explains in some detail the meaning of allocative efficiency (and productive efficiency) in the context of two market structures known as perfect competition and monopoly. Chapter 8 also considers allocative efficiency, this time in the context of market failure. Allocative efficiency is a welfare concept rather than a production concept.

QUANTITATIVE SKILLS 1.3

Worked example: drawing a *PPF* from given data

Draw a production possibility frontier using the data in Table 1.3.

Table 1.3 Production possibility schedule for producing tanks and military aircraft

Tanks	Military aircraft
100	0
90	10
80	20
70	30
60	40
50	50
40	60
30	70
20	80
10	90
0	100

The *PPF* drawn from these data is shown in Figure 1.10.

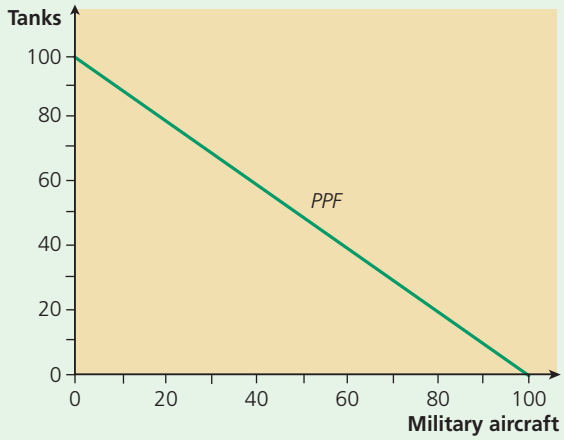


Figure 1.10 *PPF* for tanks and military aircraft

Unlike all the *PPF*s shown earlier in this chapter, this one shows a constant opportunity cost of 10 military aircraft whenever 10 extra tanks are produced. This means that the *PPF* is a straight line, positioned between the two production possibility extremes of zero tanks and 100 military aircraft, and 100 tanks and zero military aircraft.

TEST YOURSELF 1.14

Suppose the opportunity cost of producing 1 extra unit of a good in terms of another good sacrificed fell as more of the good was produced. In what way would the shape of the *PPF* differ from those shown in this chapter?

QUANTITATIVE SKILLS 1.4

Worked example: interpreting statistical data and performing simple percentage calculations

Table 1.4 shows the proportion of total output exported in selected industries, in January 2010 and January 2019.

Table 1.4 Export proportions for selected manufacturing industries in the UK

Industry	Percentage of total industry output exported, January 2010	Percentage of total industry output exported, January 2019
Alcoholic drinks	14.5	28.6
Petrochemicals	52.9	68.1
Furniture	5.5	10.0
Motor vehicles	39.5	50.6
Computers and electronic and optical products	48.2	58.7
Weapons	9.2	28.1

Source: ONS

(a) Which industry enjoyed the greatest percentage growth in the proportion of its output exported between January 2010 and January 2019?

The percentage growth in the proportion of its output exported over the year were: for alcohol drinks 97.2%; petrochemicals 28.7%; furniture 81.8%; motor vehicles 28.1%; computers etc. 21.8%; weapons 205.4%. The greatest percentage growth was in weapons exports.

(b) Is it possible to calculate the actual amount by which petrochemical output changed between January 2010 and January 2019? Explain your answer.

No. We can calculate the percentage growth but not the actual growth. We need the money value of total output of petrochemicals in January 2010 to enable this calculation to be made.

SECTION 1.5 SUMMARY

- Key economic concepts such as scarcity, choice, opportunity cost, economic growth, and full employment and unemployment, can be illustrated on a production possibility diagram.
- A production possibility frontier illustrates the different combinations of goods that can be produced with a fixed quantity of resources.
- Different economic systems allocate resources between different uses in different ways.
- In a market economy, the price mechanism performs the allocative task.
- The UK economy is a mixed economy, containing a mix of market and non-market sectors, and private and public sectors.
- The nature of the UK mixed economy has changed during the last 40 years.

Questions

1 With one example of each, explain the difference between a positive and a normative statement.

2 What is the fundamental or central economic problem?

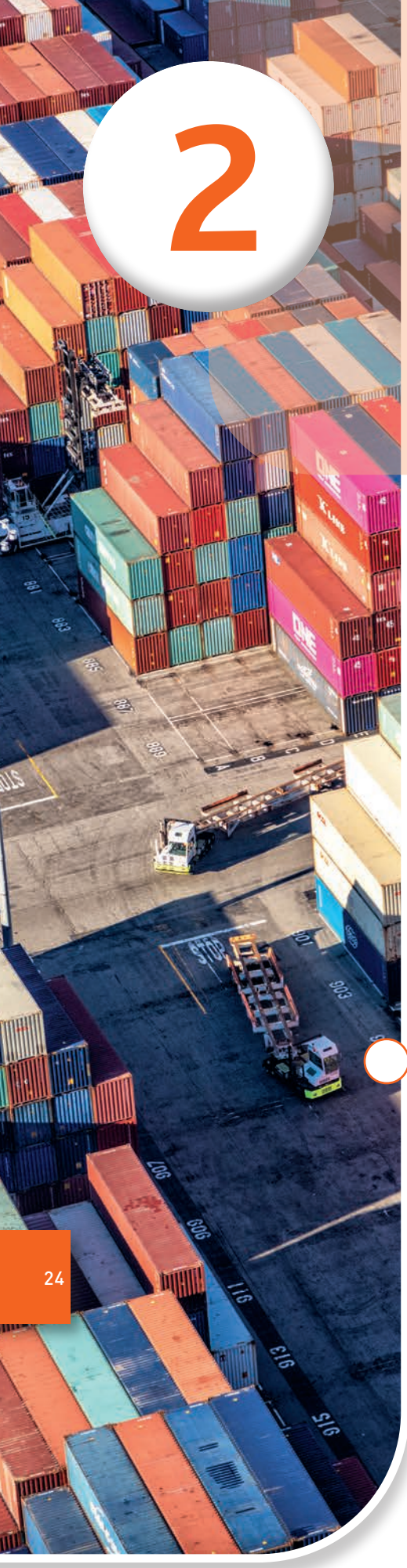
3 Relate scarcity to the need to make economic choices and to opportunity cost.

4 Briefly describe the four factors of production.

5 What is a production possibility frontier?

6 Draw a production possibility frontier to illustrate the choice between producing capital goods and consumer goods.

7 Draw a production possibility frontier which illustrates economic growth taking place.



2 Individual economic decision making

This chapter introduces two extremely significant aspects of individual economic decision making. The first is utility theory, an old-established body of theory which underlies the development of demand theory. Then, following a brief discussion of how imperfect information affects individual decision making, we introduce you to the second important part of this chapter, behavioural economics. Behavioural economics is a relatively new part of the subject which provides significant insights into how individuals make economic decisions.

LEARNING OBJECTIVES

These are to understand:

- the significance of utility maximisation for individual economic decision making
- the importance of the margin when making choices
- how imperfect information and asymmetric information affect choice decisions
- the emergence of behavioural economics as an important recent development in economic theory
- important elements of behavioural economics such as bounded rationality, biases in individual decision making and the role of altruism
- how behavioural economics relates to government economic policy.

2.1 Consumer behaviour

Rational economic decision making and economic incentives

At the heart of traditional or orthodox demand theory is the assumption that the members of households or consumers always act rationally. **Rational behaviour** means people try to make decisions in their self-interest or to maximise their private benefit. When a choice has to be made, people always choose what they think at the time is the best alternative, which means that the second best or next best alternative is rejected. For households and the individuals within them, rational behaviour means attempting to maximise the welfare, satisfaction or utility gained from the goods and services consumed.

KEY TERM

rational behaviour acting in pursuit of self-interest, which for a consumer means attempting to maximise the welfare, satisfaction or utility gained from the goods and services consumed.

SYNOPTIC LINK

Chapter 3 will look at how traditional economic theory explains price determination. At the heart of this approach is the assumption that individuals make rational decisions and seek to maximise utility.

TEST YOURSELF 2.1

Which one of the following provides the best reason why consumers become early adopters of a new innovation such as a smart watch, even though they know that they will be paying a high price for the good?

Early adopters are people who:

A like technological gadgets

B get up early to buy in a sale

C base their consumption decisions on the reviews submitted online by existing users

D want to be the first to get new types of product as they come onto the market

Explain your answer.

Utility theory: total and marginal utility, and diminishing marginal utility

What is utility?

Consumers attempt to maximise the welfare or utility they gain from the goods and services they decide to consume. We will explore this further in the next section, on utility maximisation. In economics, **utility** is usually defined as the pleasure or satisfaction obtained from consumption, and **marginal utility** is the *additional* pleasure or satisfaction obtained from consuming one more unit of something. Although it is difficult to compare the utility obtained by each individual, we can assign nominal values to different levels of utility.

KEY TERMS

utility the satisfaction or economic welfare an individual gains from consuming a good or service.

marginal utility the additional welfare, satisfaction or pleasure gained from consuming one extra unit of a good or service.

TEST YOURSELF 2.2

A family's typical weekly shopping basket might include 'pleasure items' such as packets of crisps and cans of Coca-Cola and other items which fulfil a need and without which life would be more uncomfortable. Are each of the following 'pleasure items' or 'need-fulfilment items'?

(a) medicine

(b) chocolate

(c) daffodil bulbs

(d) electric light bulbs

(e) washing-up liquid

The relationship between total utility and marginal utility

Let us imagine a thirsty teenager who drinks 6 glasses of lemonade on a hot sunny afternoon, deriving successively 8, 6, 4, 2, 0 and –2 'units of utility' from each glass consumed. This information is shown in the total and marginal utility schedules in Table 2.1, from which the total and marginal utility curves in Figure 2.1 are plotted. Note that marginal utility is plotted at 'halfway' points.

Table 2.1 Total and marginal utility schedules for lemonade

Glasses of lemonade	Total utility (units of utility)	Marginal utility (units of utility)
0	0	–
1	8	8
2	14	6
3	18	4
4	20	2
5	20	0
6	18	–2

It is important to realise that the total and marginal utility schedules and, likewise, the total and marginal utility curves show exactly the same information, but they show it in different ways. The total utility schedule and the total utility curve show the data cumulatively — for example, when drinking two glasses of lemonade, the thirsty teenager gains 14 'units of utility' in total. After three glasses, total utility rises to 18 'units of utility', and so on.

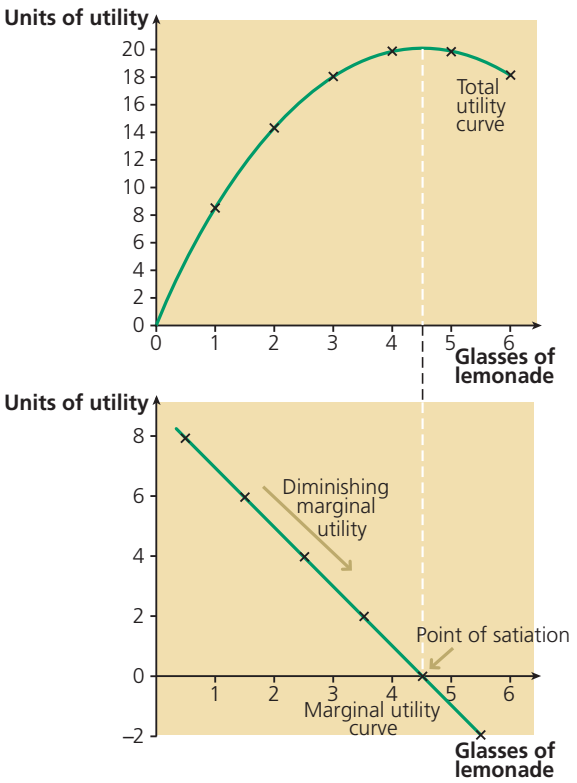


Figure 2.1 An example of total utility and marginal utility curves

In contrast, the marginal utility schedule and the marginal utility curve plot the same data as separate observations, rather than cumulatively. The last unit consumed is always the marginal unit and the utility derived from it is the marginal utility. So, after two drinks, the second glass of lemonade is the marginal unit consumed, yielding a marginal utility of 6 'units of utility'. But when three glasses of lemonade are consumed, the third glass becomes the marginal unit, from which the still partially thirsty teenager gains a marginal utility of just 4 'units of utility'.

In Figure 2.1, diminishing marginal utility is shown both by the diminishing rate of increase of the slope of the total utility curve drawn in the upper panel of the diagram and by the negative or downward slope of the marginal utility curve in the lower panel. Notice that we have drawn a 'point of satiation' on the diagram, which is reached as the fifth glass of lemonade is drunk. The fifth glass of lemonade yields zero marginal utility. At this point, when marginal utility is zero, total utility is maximised. In the context of food and drink, satiation means being 'full up'. Even if lemonade is free to the consumer, it would be irrational for our teenager who is no longer thirsty to drink a sixth glass of lemonade. He or she would experience negative marginal utility (or marginal disutility), which is shown by the downward slope of the total utility curve and by the negative position of the lower section of the marginal utility curve.

QUANTITATIVE SKILLS 2.1

Worked example: calculating marginal utility

An 8-year-old boy decides to enter a competition to see how many jam doughnuts can be eaten in 15 minutes. Table 2.2 shows how many he ate and his total utility schedule.

Table 2.2 Total utility for doughnuts

Jam doughnuts	Total utility (units of utility)	Marginal utility (units of utility)
0	0	
1	6	
2	10	
3	12	
4	12	
5	8	
6	3	

Complete the boy's marginal utility schedule.

The boy's marginal utility schedule is shown in Table 2.3.

Table 2.3 Total and marginal utility for doughnuts

Jam doughnuts	Total utility (units of utility)	Marginal utility (units of utility)
0	0	–
1	6	6
2	10	4
3	12	2
4	12	0
5	8	−4
6	3	−5

KEY TERM

hypothesis of diminishing marginal utility for a single consumer, the marginal utility derived from a good or service diminishes for each additional unit consumed.

The hypothesis (or 'law') of diminishing marginal utility

The numerical examples in Tables 2.1–2.3, and the graph in Figure 2.1, illustrate a famous economic hypothesis, which some would call an economic law: the **hypothesis of diminishing marginal utility**. This simply states that as a person increases consumption of a good — while keeping consumption of other products constant — there is a decline in the marginal utility derived from consuming each additional unit of the good.

TEST YOURSELF 2.3

Explain how traditional economic theory assumes a rational individual will make decisions: for example, when deciding how to respond to a change in a good's price.

SYNOPTIC LINK

In the context of economic methodology, Chapter 1 explained the difference between a hypothesis and a theory. To remind you, whereas a hypothesis is a proposed explanation for something, a theory is when a hypothesis is tested and survives the test.

APPLICATION OF ECONOMICS IN THE REAL WORLD 2.1

Adam Smith's diamonds and water paradox

In 1776 the great classical economist Adam Smith wrote about the diamonds and water paradox (or the paradox of value) in his famous book *The Wealth of Nations*. Smith wrote:

> Nothing is more useful than water: but; scarce any thing can be had in exchange for it. A diamond, on the contrary, has scarce any value in use; but a very great quantity of other goods may frequently be had in exchange for it.

In most countries, water has a low price, but a piece of diamond jewellery has a high price. Why does an economy put a much lower value on something vital to sustaining life compared to something that simply looks good? Smith pointed out that practical things that we use every day have a *value in use*, but often have little or no *value in exchange*. On the other hand, some of the things that often have the greatest value in the market or in exchange, such as a drawing by Picasso, have little or no practical use other than, in this case, as ornamentation.

Understanding the diamonds and water paradox comes through first understanding the economic terms 'marginal utility' and 'scarcity'. Scarcity relates to how little of a good there is compared to what people are demanding. Marginal utility is the additional welfare a person gains from using or purchasing an additional unit of the good. People are willing to pay a higher price for goods with greater marginal utility.

Relating this to water and diamonds, water is not scarce in most of the world, which means people can consume water up to the point at which the marginal utility they gain from the last drop consumed is very low. They aren't willing to pay a lot of money for one more drink of water. Diamonds, by contrast, are scarce. Because of their limited supply, the marginal utility typically gained from adding one more diamond to a person's collection is much higher than for one extra drink of water. However, if one is dying of thirst, this paradox breaks down. In this situation, the marginal utility gained from another drink of water would be much higher than the additional satisfaction of owning an extra diamond — at least until the thirst was quenched.

Follow-up questions

1. Define the terms 'scarcity' and 'marginal utility'.
2. Can you think of two other goods which generally illustrate the paradox of value?
3. Why are a bottle of water and a diamond priced differently?
4. Is it rational for an individual to trade a diamond for a bottle of water?

SYNOPTIC LINK

This chapter focuses on the individual choices made by consumers or members of households when they decide how much of a good to consume in the economy's goods or product market. In Chapter 6, by contrast, we explain the choices made by members of the same households about how much labour to supply and how much leisure time to enjoy when making decisions in the economy's labour market. In both sets of markets, utility maximisation (explained below) is central to individual economic decision making. In the labour market we assume that workers attempt to maximise the utility derived from the wage and the utility gained from job satisfaction.

Utility maximisation

The assumption of maximising behaviour by economic agents (consumers, workers, firms and even the government) is central to traditional economic theory. Economic agents decide their market plans so as to maximise a target objective or goal which is believed to be consistent with the pursuit of self-interest. In demand theory, the objective which households are assumed to wish to maximise is the utility, or satisfaction, obtained from the set of goods and services consumed.

EXTENSION MATERIAL

Maximising versus minimising behaviour

It is worth noting that any maximising objective can always be recast as a minimising objective. Thus a household's assumed objective of *maximising the utility gained from the set of goods and services consumed* can be restated as *minimising the outlay, expenditure or cost of obtaining the same combination or bundle of goods and services*. Whether we set up an assumed objective in maximising or minimising terms depends on our convenience. It is more usual to investigate maximising objectives, but for some purposes a consideration of the minimising principle can shed interesting light on economic behaviour.

EXTENSION MATERIAL

Maximisation subject to constraints

If all goods were free, or if households had unlimited income and capacity to consume all goods, a consumer would maximise utility by obtaining all the goods which yield utility, up to the point of satiation. As we have already indicated, satiation occurs when no more utility can be gained from consuming extra units of a good. Any further consumption would yield only disutility at the margin (negative utility, dissatisfaction or displeasure).

However, because of the problem of scarcity, consumers face a number of constraints which restrict the choices they make in the marketplace. The constraints are:

- **Limited income.** Consumers, even the very rich, do not possess an unlimited income, or stock of wealth that can be converted into income, with which to purchase all the goods and services that could possibly yield utility. Income spent on one good cannot be spent on some other good or service.
- **A given set of prices.** Very often, consumers cannot by their own actions influence the market prices they have to pay to obtain the goods and services they buy. Given this assumption, consumers are 'price-takers' rather than 'price-makers'.
- **The budget constraint.** Taken together, limited income and the set of prices faced impose a budget constraint on consumers' freedom of action in the marketplace. If we assume that all income is spent and not saved, that there is no borrowing, and that stocks of wealth are not run down, a consumer can only purchase more of one good by giving up consumption of some other good or service, which represents the opportunity cost of consumption.
- **Limited time available.** Even when goods are free, consumer choices must still be made because it is often impossible to consume more than one good at a time or to store more than a limited number of goods for future consumption.

Importance of the margin when making choices

Along with assumptions such as rational economic behaviour and opportunity cost, the 'margin' is one of the key concepts in traditional economic theory. Given consistent tastes and preferences, rational consumers choose between available goods and services in such a way as to try to maximise total utility, welfare or satisfaction derived from consumption of the goods. Along with the relative prices that must be paid for each of the goods, the marginal utilities gained from the consumption of the last unit of each good determine the combination of goods the consumer must choose in order to maximise total utility.

As we will see in later chapters in Part 1 of this book, in traditional economic theory, the margin is equally important in other areas of economic choice. For example, we will see how, when firms choose how much of a good to produce and sell, they take account of the marginal sales revenue received from selling the last unit of the good, and the marginal cost of producing the last unit. Generalising across all choice situations, we will explain how, in order to maximise a desired objective, an economic agent must undertake the activity involved up to the point at which the marginal private benefit received equals the marginal private cost incurred. For example, a utility-maximising consumer must choose to consume or demand a good up to the point at which $MU = P$. Marginal utility or MU is, of course, the marginal private benefit derived from consuming the last unit of the good, while the good's price, P, is its opportunity cost in consumption, at the margin.

STUDY TIP

The margin is one of the key concepts in A-level microeconomics. You have already encountered the concept of marginal utility in this chapter. In Chapters 4, 5 and 6 you will come across concepts such as marginal cost, marginal revenue and marginal returns. Make sure you understand and can define 'the margin' and apply the concept in its different forms in your exam.

EXTENSION MATERIAL

Can utility be measured?

On several occasions we have referred to 'units of utility' as a unit of measurement for the happiness, pleasure, satisfaction or fulfilment of need which an individual derives from consuming a good or service. However, in real life there is no way in which an individual can mathematically work out the utility to be gained from every unit of a good consumed. Economists have found it impossible to measure directly units of satisfaction, pleasure or fulfilment through which comparisons can be made across individuals.

To get around this problem, the famous economist Paul Samuelson introduced the concept of 'revealed preference'. What revealed preference theory does is work backwards from observing how consumers actually behave to observing their preferences. Consumers reveal their preferences by choosing, at given prices and for given levels of income, the bundles or combinations of goods they end up buying.

SECTION 2.1 SUMMARY

- The starting point for understanding individual economic decision making is understanding the nature of rationality and maximising behaviour.
- Economists have traditionally assumed that individuals wish to maximise utility.
- Utility can be thought of as satisfaction, pleasure or fulfilment of need.
- It is important to distinguish between total utility and marginal utility.
- The hypothesis (or 'law') of diminishing marginal utility lies behind the derivation of an individual's demand curve.
- Utility cannot be measured directly.

2.2 Imperfect information

The importance of information for decision making

So far in this chapter, we have assumed that consumers possess perfect information: for example, about the goods that are available to buy, their prices and quality, and about the utility which will be derived from their consumption. Traditional economic theory assumes that if individuals possess perfect information, they will make decisions that maximise their welfare. However, when attempting to maximise total utility, more often than not consumers possess imperfect information. As a result, they make 'wrong' decisions.

A student may spend £100 on a ticket for a rock concert, believing in advance that she will thoroughly enjoy the entertainment. However, she may come out of the stadium in which the event was held believing that she had wasted her hard-earned money and would be far better off if she had spent the £100 on other goods, such as a meal in an expensive restaurant. This is an example of a 'wrong' choice, but it was also a rational choice because she believed in advance that the concert would be good.

In Chapter 8, section 8. 5, we will study how consumers may choose to under-consume a merit good such as education and over-consume a demerit good such as tobacco because they possess imperfect information about the long-term consequences of their choices. We will investigate the contribution of behavioural economics in the next section of this chapter.

SYNOPTIC LINK
In Chapter 8 we analyse merit goods and demerit goods by examining the information failures that households experience when deciding how much to consume of a merit good such as education, or a demerit good such as tobacco.

The significance of asymmetric information

Sometimes, one party to a market transaction, either the buyer or the seller, suffers from imperfect information about the nature of the transaction. **Asymmetric information** arises when either the buyer or the seller involved in a potential transaction knows something that is not observable to the other party. One of the ways in which asymmetric information can manifest itself is through the process known as *adverse selection*, which is a feature of many market transactions. For example, in the sale and purchase of a second-hand computer, the seller of the good knows more about the computer's defects than a potential purchaser. However, to avoid paying too high a price for an inferior product which has lots of defects, potential purchasers often offer low prices on *all* second-hand computers, regardless of the fact that some of the computers are good.

KEY TERM
asymmetric information
when one party to a market transaction possesses less information relevant to the exchange than the other.

The problem of asymmetric information possessed by buyers and sellers is described in a classic article by George Akerlof on the market for 'lemons' — a 'lemon' being American slang for a poor-quality second-hand car.

TEST YOURSELF 2.4
Why does asymmetric information give a salesperson an advantage in the market for used cars or in the market for private pensions?

APPLICATION OF ECONOMICS IN THE REAL WORLD 2.2

The market for 'lemons'

In 2001 George Akerlof was awarded the Nobel prize in economics, largely in response to a 13-page academic paper he published in 1970 titled 'The market for lemons'. Back in 1970, Akerlof found it difficult to get his paper published. Two leading academic journals rejected it on the grounds that asymmetric information in the market for second-hand cars was too trivial an economic issue. However, by 2001 things had changed.

On receiving his Nobel prize, Akerlof said:

> 'Lemons' deals with a problem as old as markets themselves. It concerns how horse traders respond to the natural question: 'if he wants to sell that horse, do I really want to buy it?' Such questioning is fundamental to the market for horses and used cars, but it is also at least minimally present in many market transactions.

Here is an extract from what Akerlof wrote in his 1970 paper:

> From time to time one hears either mention of or surprise at the large price difference between new cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a 'new' car.
>
> We offer a different explanation. Suppose that there are just four kinds of cars. There are new cars and used cars. There are good cars and bad cars (which in America are known as 'lemons'). A new car may be a good car or a 'lemon', and of course the same is true of used cars.
>
> The individuals in this market buy a new automobile without knowing whether the car they buy will be good or a 'lemon'. After owning a specific car, however, for a length of time, the car owner can form a good idea of the quality of this machine.
>
> An asymmetry of information has developed: for the sellers have more knowledge about the quality of a car than the buyers. But good and bad used cars must still sell at the same price, since it is impossible for a buyer to tell the difference between a good and a bad car.
>
> It is apparent that a used car cannot have the same valuation as a new car — if it did, it would clearly be advantageous to trade a lemon at the price of a new car, at a high probability of the new car being a good car. Most used cars traded will be 'lemons', and good used cars may not be traded at all. The 'bad cars tend to drive out the good (in much the same way that bad money drives out the good)'.

Follow-up questions

1. Define the term 'asymmetric information'.
2. Explain why a buyer with a good knowledge of car mechanics has an advantage in the market for second-hand cars.
3. Explain, using examples other than cars, which markets are most affected by the problem of asymmetric information.
4. Discuss to what extent you believe that, if the price of a good is too good to be true, you are buying a 'lemon' rather than taking advantage of an opportunity.

SECTION 2.2 SUMMARY

- Traditional economic theory assumes individuals will make rational choices.
- In reality, individuals have to make decisions based on imperfect information, which makes them prone to making a 'wrong decision'.
- In real-world markets, the problem of asymmetric information arises when either the buyer or the seller knows something not observable to the other party.

2.3 Aspects of behavioural economic theory

Emergence of behavioural economics

KEY TERM
behavioural economics a method of economic analysis that applies psychological insights into human behaviour to explain how individuals make choices and decisions.

STUDY TIP
Try to read Thaler and Sunstein's *Nudge: Improving Decisions about Health, Wealth and Happiness*, and also some of the Behavioural Insight Team's publications, which can be accessed on the internet. Both will give you a deeper understanding of how behavioural economics can help to encourage choices intended to increase the overall welfare of society. This will help you to address questions about how governments could correct instances of market failure — for example, presenting meal options in school canteens in ways which promote healthier choices.

Behavioural economics is a field of study that has attracted a great deal of attention since the beginning of the twenty-first century. Most of the research in the field has come from universities in the USA, but in recent years UK university economics departments have been offering courses in the subject. Behavioural economics is built on the insights of psychologists seeking to understand human behaviour and decision making. In 2008, economist Richard Thaler and legal scholar Cass Sunstein published *Nudge: Improving Decisions about Health, Wealth and Happiness*, which is a highly accessible overview of behavioural economics. In 2010, the UK government set up the Behavioural Insights Team (BIT), which was initially based in the Cabinet Office in Downing Street. The creation of the BIT, and of a similar body advising the then US president, Barack Obama, marked the growing influence that behavioural economics was having on government policy-makers.

On its website, the BIT writes:

> We coined the term 'behavioural insights' in 2010 to help bring together ideas from a range of inter-related academic disciplines (behavioural economics, psychology, and social anthropology). These fields seek to understand how individuals take decisions in practice and how they are likely to respond to options. Their insights enable us to design policies or interventions that can encourage, support and enable people to make better choices for themselves and society.

The BIT's website is worth reading to find out about projects that the team runs in government. You might also access a BIT publication, *Better Choices: Better Deals*, otherwise known as the UK government's consumer empowerment strategy, which recommends how government policy can attempt to influence consumer behaviour. It can be found at **www.gov.uk/government/publications/better-choices-better-deals-behavioural-insights-team-paper**.

TEST YOURSELF 2.5
Which one of the following statements is true about how the use of the internet affects consumer behaviour?

A It facilitates the use of price comparison websites.

B It stops consumers comparing prices in stores.

C It always adds to consumer confusion.

D It eliminates the need for consumer choice.

Explain your answer.

EXTENSION MATERIAL

Squaring the circle between traditional and behavioural economic theory

In his excellent book *Predictably Irrational*, Dan Ariely stated that traditional economics is about creating a theory and using it to explain actual behaviour, whereas behavioural economics is about observing actual behaviour and then coming up with a theory.

Traditional theories are often attacked by behavioural economists on the grounds that the simplifying assumptions on which the theories are built are unrealistic. In particular, in the context of what traditional economists call the 'theory of the firm', behavioural economists query the 'profit-maximising assumption'. This is the assumption that entrepreneurs make business decisions solely on the basis of whether the decisions will lead to larger profits.

However, in the essay, 'The methodology of positive economics', published in 1953, the free-market economist Milton Friedman defended the traditional approach. Friedman wrote: 'Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions.'

Friedman rejected testing a theory solely on the realism of its assumptions. He agreed that assumptions such as utility maximisation and profit maximisation are unrealistic. Friedman argued that a theory should be tested and then accepted or rejected on the basis of the validity and fruitfulness of its predictions. If unrealistic assumptions led to wrong conclusions, he would have argued that the theory should be rejected or modified. But if assumptions are unrealistic because of the need to abstract from a complex reality, but still lead to sound predictions which survive scientific testing, they can be justified. In summary, if members of households act 'as if' they are utility maximisers and likewise the entrepreneurs who run firms act 'as if' they are profit maximisers, the predictive power of traditional theories can still be good.

In the traditional theory of the firm, entrepreneurs are assumed to produce and sell output up to the point at which marginal revenue equals marginal cost, yet real-world business people seldom make such decisions when running their businesses. Friedman argued that this does not matter. If Friedman had lived to the present day (he died at the age of 94 in 2006), he might be using similar reasoning to defend traditional economic theory from the attacks of behavioural economists.

SYNOPTIC LINK

The traditional theory of production, known as the theory of the firm, assumes that rational producers always seek to maximise profit as their reward for taking a risk in starting a business. The traditional theory of the firm and profit maximisation are explained in depth in Chapters 4 and 5.

STUDY TIP

Make sure you understand the key differences between traditional economic theory and behavioural economics. Each of these makes important assumptions about how individuals and firms behave. Note too, however, that each set of theories can complement the other. This is an important consideration when tackling questions that ask about the 'best' methods of tackling economic problems. A good approach is to explain how more mainstream policies may be augmented by those which take into account an appreciation of behavioural economics.

APPLICATION OF ECONOMICS IN THE REAL WORLD 2.3

In the April 2010 edition of the Economic Review, *published by Philip Allan for Hodder Education, David Gill presented an overview of interesting developments in behavioural economics. This case study summarises the introduction of David Gill's article.*

Beyond *Homo economicus*

Economists like to simplify the world; in particular they like to simplify people. Most of twentieth-century economics makes a number of standard assumptions about how people behave, which comprise our view of *Homo economicus* or 'economic man'. *Homo economicus* is self-interested: he only cares about himself. He knows the consequences of everything he does. He is rational: he knows what he wants and always acts on these preferences.

This simple model has proved to be exceptionally useful in gaining insights into economic behaviour, especially when consumers and firms interact in large markets. However, the new science of behavioural economics seeks to move beyond *Homo economicus* to a more realistic representation of how people choose and behave. It does so in a number of ways.

First, data collected by economists show a number of so-called 'anomalies' — that is, behaviour which deviates in a consistent manner from that predicted by the model of *Homo economicus*. For example, people appear to be altruistic: they tend to put at least some weight on the well-being of others. Another point is that we are generally impatient and lack self-control. Most of us find it difficult to resist immediate temptation, whether it be a chocolate bar or an extra hour in bed. Finally, we dislike change. The so-called 'status quo bias' means that we generally like to stick with what we have, unless the incentive to change course is compelling.

Second, research from the field of cognitive psychology paints a picture at odds with that of *Homo economicus*. This research shows that humans often make decisions using simple rules-of-thumb — called *heuristics* — and suffer from many biases when choosing what to do, such as over-confidence, *confirmation bias* (the tendency to search for, and put greater weight on, information that confirms one's preconceptions) and *recency bias* (the tendency to weight recent information and experience more heavily than older information and earlier experiences). Also psychological findings emphasise the fundamental role of emotions in decision making, including, for example, anger, regret, guilt, shame and disappointment.

Third, humans clearly face quite substantial limitations of computation and reasoning. These are particularly important when the environment is complex: for example, when people are interacting strategically, my best action will depend on what others choose to do. In production, firms in markets known as oligopolies, when deciding what price to set, will have to start thinking about what their competitors are going to do.

Follow-up questions

1. How does 'economic man' make decisions?
2. Define the term 'altruism'.
3. Is it rational for an individual to give money to charity or to donate blood?
4. To what extent is the UK mortgage market a complex market?

SYNOPTIC LINK

Oligopolistic pricing behaviour is explained in Chapter 5, pages 145–50. This gives several explanations of how firms in concentrated markets pursue an optimal strategy to reduce uncertainty.

Bounded rationality

So far in this chapter, we have assumed that when exercising choice, individuals are perfectly rational, in the sense that they make decisions in a context of being fully informed, with perfect logic and aiming to achieve the maximum possible economic gain. However, in real life, individuals are seldom if ever perfectly rational.

KEY TERM
bounded rationality when making decisions, individuals' rationality is limited by the information they have, the limitations of their minds, and the finite amount of time available in which to make decisions.

In the world in which we live, decisions are made in conditions of **bounded rationality**, which means that individuals, however clever they are, make decisions subject to three unavoidable constraints: imperfect information about possible alternatives and their consequences; limited mental processing ability; and a time constraint which limits the time available for making decisions. In complex choice situations, bounded rationality often results in *satisficing* rather than *maximising* choices. Satisficing means accepting outcomes which are sub-optimal, or second-best.

SYNOPTIC LINK
The difference between firms profit satisficing and maximising is explained in Chapter 5, pages 127–29. In practice, although firms may wish to maximise their profits, there may be several reasons why they are unable to. Therefore, they may settle for a sub-optimal, satisficing level of profit.

Bounded self-control

KEY TERM
bounded self-control where individuals have limited self-control to act rationally in their own interests.

Bounded rationality is closely linked to the related concept of **bounded self-control**. Traditional economic theory implicitly assumes that when making choices, individuals have complete self-control. Behavioural economists, by contrast, believe that some individuals have bounded (or limited) self-control. Such people may set out with good intentions to act rationally, but find themselves slipping into less rational behaviour as more appealing alternatives are encountered. For example, a healthy-eating regime may give way to fast food after a few days.

Thinking fast and thinking slow

The Nobel prize-winning psychologist Daniel Kahneman has been one of the most influential figures in the development of behavioural economics. Kahneman introduced economists to the idea that human beings think in two different ways. The first, which Kahneman called System 1 or 'thinking fast', is intuitive and instinctive. Decisions are made quickly and little effort is used to analyse the situation. This is automatic thinking.

The second, which Kahneman called System 2, is 'thinking slow'. In this method of thinking, which is also known as reflective thinking, concentration and mental effort are required to work through a problem before a decision can be made.

For example, when learning a new skill such as skateboarding, an individual will 'think slow' when practising the appropriate skill. Because the decision making is relatively slow, involving careful, logical thought about every decision, the process can be tiring. However, the more often the skills are practised, the less skateborders have to think about minor decisions. Automatic thinking takes over. Skateboarding experts often act quickly and instinctively. Through years of repetitive training, their automatic systems have learnt to respond to situations promptly and effectively. In high-pressure situations they can, of course, suffer if they stop to think. When this happens, they switch back to System 2 or the reflective system, over-thinking the situation, which can mean that bad decisions have disastrous consequences.

Many of our everyday economic decisions are taken by our automatic system. Buying a coffee at a train station, buying groceries in a supermarket and buying a sandwich from a snack bar will often be quick, intuitive decisions. Bigger and more important decisions tend to be taken by our reflective system. Deciding whether to buy a car or a house, and choosing an insurance policy, will normally result from reflective decisions.

Biases in decision making

Behavioural economics argues that quick decisions that people make when exercising choice (automatic thinking) are often heavily biased. This is because decisions are made on the basis of one's own likes, dislikes and past experiences. Psychologists use the term **cognitive bias** to describe this situation. A cognitive bias is a mistake in reasoning or other thought process, often occurring as a result of holding onto one's preferences and beliefs regardless of contrary information. Many kinds of cognitive bias exist, such as confirmation bias, which is the tendency to seek only information that matches what one already believes.

KEY TERMS

cognitive bias a systematic error in thinking that affects the decisions and judgements that people make.

rules of thumb thinking shortcuts, or informed guesses, that individuals use to make decisions in order to save time and effort.

availability bias occurs when individuals make judgements about the likelihood of future events according to how easy it is to recall examples of similar events.

anchoring a cognitive bias describing the human tendency when making decisions to rely too heavily on the first piece of information offered (the so-called 'anchor'). Individuals use an initial piece of information when making subsequent judgements.

Rules of thumb

Rules of thumb are thinking shortcuts used by humans to help them make sensible decisions based on limited information, but the decisions are prone to cognitive bias. The two common rules of thumb that lead to biases are availability and anchoring.

The availability bias

The **availability bias** occurs when individuals place too much weight on the probability of an event happening because they can recall vivid examples of similar events. For example, after reading several news reports about car thefts, an individual may judge that vehicle theft is much more common than it really is in the local area.

The availability bias often leads to decisions that are not based on logical reasoning. The media report stories that stick in our mind and affect our reasoning process. Humans often believe that the probability of an extreme weather event, such as a hurricane or severe flooding, is more likely than empirical statistical analysis bears out. The market research company Ipsos has published research highlighting how the general public in 14 countries held preconceptions on the make-up of their societies that were significantly detached from the reality.

In the UK, for example, many citizens believed that 24% of the population were immigrants when the real figure was 13% and, likewise, that 24% of the working age population was unemployed when in fact at the time of the research it was less than 7%.

Anchoring

Anchoring is an example of a predictable bias in individual decision making. Most people have a tendency to compare and contrast only a limited set of items. This is called the anchoring effect. A good example is provided by restaurant menus, which sometimes feature very expensive main courses, while also including more (apparently) reasonably priced alternatives. We are lured into choosing the cheaper items, even though their prices are still quite high. When given a choice, we often tend to pick the middle option, believing it is not too expensive, but also not too cheap.

TEST YOURSELF 2.6
Explain how individuals use 'rules of thumb' when making decisions.

KEY TERMS
social norms forms or patterns of behaviour considered acceptable by a society or group within that society.

nudges factors which encourage people to think and act in particular ways. Nudges try to shift group and individual behaviour in ways which comply with desirable social norms.

TEST YOURSELF 2.7
Explain how a social norm can lead to less litter and chewing gum being dropped on a street.

Biases based on social norms

Human beings are social animals and as a result the behaviour of other people influences our own behaviour. By unconsciously learning from the behaviour of other people, **social norms** are established.

Negative social norms include attitudes towards drinking alcohol. Many young adults often drink heavily because they think it is what people of their age are expected to do. By presenting statistical data showing that the majority of young adults don't engage in regular heavy drinking, behavioural economists would seek to *nudge* young drinkers into different patterns of behaviour.

Positive social norms can be seen in the way in which social attitudes have altered towards smoking in the last 30 years. In the 1980s it was socially acceptable to smoke in all public places including libraries, trains and the London Underground. Concerted health campaigns, which provided the general public with better information about the risks of smoking, have altered social attitudes. As a result, people became much more willing to accept laws which restricted their right to smoke. The laws banning smoking in public places are economic sanctions (used by government policy-makers) and not **nudges**. Economic sanctions include restrictions imposed by regulations and/or laws that restrict an individual's freedom to behave in certain ways. Breaking a sanction can lead to punishment.

Critics of behavioural economics point out that sanctions, such as the smoking ban, may be more effective at changing behaviour and improving public health than nudges, which only alter the behaviour of some people. Nevertheless, government reports in Ireland claim that since smoking in public places was banned, people are also less likely to smoke in other people's houses because it is now considered socially unacceptable.



The economist Esther Duflo has won a Nobel prize for, among other things, her work on using social aspects of behavioural economics to help reduce poverty and improve economic development. One example, from 2003, involved reducing teacher absenteeism in India by encouraging teachers in 120 schools to have their photograph taken with their students each day in order to reinforce the bond between teacher and learner.

French-American economist Esther Duflo has used behavioural economics to improve economic development in India

SYNOPTIC LINK

Nudges are a good example of how policies that take behavioural economics into account can complement 'traditional' policies, such as indirect taxation, to correct market failures. See Chapter 8 for a discussion of nudges in relation to environmental market failures.

TEST YOURSELF 2.8

Which one of the following provides the best definition of norms?

Norms are:

A laws that attempt to discourage excessive consumption

B informal rules that govern human behaviour

C formal rules that govern human behaviour

D formal rules about how to buy goods and services

Explain your answer.

Altruism and fairness

KEY TERMS

altruism concern for the welfare of others.

fairness the quality of being impartial, just, or free of favouritism. It can mean treating people equally, sharing with others, giving others respect and time, and not taking advantage of them.

Altruism is when we act to promote someone else's well-being, even though we may suffer as a consequence, either in terms of a financial or time loss, or by incurring personal risk. Before the development of behavioural economics, economists generally assumed that individuals were not altruistic and acted only in their self-interest. Nevertheless, altruism could still be accommodated within maximising theory — for example, by assuming that individuals derive pleasure as a result of giving to others. More recently, behavioural economists have drawn attention to the fact that, for many if not most people, their first impulse is to cooperate with each other rather than to compete. Very young children are frequently observed helping other children around them, out of a genuine concern for their welfare. Animals have also been observed displaying altruism.

Altruistic behaviour often results from people's perceptions of **fairness**. Fairness is a normative term incorporating value judgements, as different people have a range of different views on the meaning of fairness. A popular view is that fairness involves treating people equally or in a way that is right or reasonable.

SYNOPTIC LINK

Go back to Chapter 1 to remind yourself of the meaning of normative statements. The notion of fairness is a normative issue because different societies will have a range of opinions about, for example, what degree of inequality is fair.

SECTION 2.3 SUMMARY

- In recent years, behavioural economics has emerged to question many of the assumptions of traditional economic theory.
- Behavioural economists believe that individuals in the real world have bounded rationality.
- An individual's economic decision making may be biased.
- Humans use rules of thumb to help them make decisions based on limited information.
- The two most common rules of thumb that lead to bias are availability and anchoring.
- Social norms can lead people into positive and negative behaviour.
- People are often altruistic and instinctively cooperate rather than compete.
- Most individuals believe in fairness and are concerned with the welfare of others.

2.4 Behavioural economics and economic policy

As we mentioned at the beginning of section 2.3, UK and US governments have recently been introducing the insights of behavioural economics into practical policy making. In assessing the impact of behavioural economics on government economic policy making, you need to consider how it might influence the design of a variety of government policies which aim to reduce or eliminate particular economic problems.

At times in this chapter, we have tended to portray traditional economics and behavioural economics as if they are completely opposed to each other, implying that if one is correct, the other is inevitably wrong. However, this is a somewhat misguided way of viewing the two very important branches of economic theory. It is better to think of behavioural economics as complementing and improving traditional economic theory by allowing governments and decision-makers to design policy interventions, such as healthcare interventions, to enable them to achieve policy goals more effectively.

As we have seen, behavioural economics argues that individuals are not fully rational in the way traditional economic theory assumes. As a result, they regularly suffer from behavioural biases that make it difficult for them to achieve the results they actually prefer. In this situation, government intervention should aim at helping individuals to achieve an outcome that is in their own best interest.

Choice architecture and framing

Choice architecture

Choice architecture is the term used by behavioural economists to describe how government policy-makers can guide people into making better choices. Government can use behavioural insights to design choice architectures so that citizens are *nudged* to opt for choices that are deemed to be in their best interest, so as to achieve a socially desirable outcome. For example, the layout and design of a canteen affects the menu choices of some individuals when purchasing a meal. Behavioural studies have found that when fruit and salads are placed at the front of a canteen more are bought than when they are placed on the back counters. Consumers are free to choose their meals but the choice architecture nudges individuals into buying healthier food.

TEST YOURSELF 2.9
Explain how well-designed choice architecture can help to solve the problem of imperfect information: for example, when designing the layout of a school canteen to encourage the eating of healthier foods.

KEY TERMS
choice architecture a framework setting out different ways in which choices can be presented to consumers, and the impact of that presentation on consumer decision making.

default choice an option that is selected automatically unless an alternative is specified.

SYNOPTIC LINK
You can find details of automatic pension enrolment in Chapter 7, page 221. This is a good example of how governments have used choice architecture to encourage individuals to make desirable choices about saving for the future.

Default choice

The **default choice** is a key aspect of choice architecture in which one option is automatically selected unless an alternative is specified. For example, when framing policy on the issue of organ donation, individuals who might donate body organs such as hearts or livers in the event of their death can be asked whether to



Signing up to the NHS Organ Donor Register is an opt-out default choice

'opt in' or 'opt out' of organ donation. In this context, an 'opt-in' default choice is illustrated by the use of a tick box which, if filled in by a member of the general public, indicates positively that they would like to donate their body organs after their death. Unless the user ticks the box, however, healthcare organisations cannot make use of their organs. This is in contrast with an 'opt-out', where the default position is that the organs can be used to help others survive unless a box has been ticked to indicate that the body parts should *not* be used. England adopted an opt-out system in 2020. Countries that require people to opt out of organ donations generally have a much higher proportion of the population registered to donate than countries that ask people to opt in. For a behavioural economist, the benefits of opt-out over opt-in are clear: the supply of donated organs rises to be closer to the demand for them and the nation's public health improves.

Policy-makers can improve social welfare by designing government programmes that select as a default an option that can be considered in an individual's best long-term interest. A number of examples of this approach have been trialled and introduced by the UK government's Behavioural Insights Team. One of these is automatic pension enrolment.

Framing

People are influenced by how information is presented. **Framing** is the tendency for people to be influenced by the context in which the choice is presented when making a decision. Advertisers have for many years presented consumers with choices in a manner that frames their products in a favourable light. Consider the label on a food product that reads: '90% fat-free'. Would it sell as well if the label read: '10% fat'?

KEY TERM
framing how something is presented (the 'frame') influences the choices people make.

Politicians often frame (or spin) economic statements in a manner that is favourable to the argument they are trying to make. For example, in the run-up to the EU Referendum in 2016, often referred to as 'Brexit', politicians supporting the Vote Leave campaign presented certain statistics in ways which made it appear the UK was significantly disadvantaged by its membership of the European Union. One of these was that the UK sends £350 million per week to the EU, which could have been spent on the NHS. Critics pointed out that this figure failed to take into account the substantial financial support the UK received from the EU for regional aid and agriculture.



Vote Leave's campaign bus makes a hotly contested claim

KEY TERMS

mandated choice people are required, often by law, to make a decision.

restricted choice offering people a limited number of options so that they are not overwhelmed by the complexity of the situation. If there are too many choices, people may make a poorly thought-out decision or not make any decision.

Mandated choices

A variant of default choice is **mandated choice**. This is where people are required, often by law, to make a decision. A mandated or required choice occurs when a choice architect designs a system that forces individuals to make an explicit decision and not merely go ahead with a default position.

An everyday example of a mandated choice outside of government policy is the Microsoft software installation boxes that appear on our computer screens. The Microsoft choice architects force computer users to make choices and select various options before they can move onto the next step and complete the installation process. Most people will choose the recommended settings, but they have to make an active decision to do so. Mandated choices work well with simple yes/no decisions but less well with complex decisions.

Restricted choice

Restricted choice means offering people a limited number of options, on the basis that offering too many choices is unhelpful and leads to poor decisions. Most people cannot, or cannot be bothered to, evaluate a large number of choices. The policy of requiring the energy companies to simplify their pricing structures and restrict the number of options offered to consumers is an example of 'restricted choice' in action.

Government policy-makers should consider behavioural insights when designing systems. A well-designed system should make it easier for citizens to pay for government services by setting up direct debits, using accessible language and sending text messages or emails to remind people to complete requests. Evidence from the BIT shows that personalised letters increase response rates, while asking respondents to sign forms at the top of the page and not the bottom results in more honest answers.

Revisiting nudge theory

As explained earlier, a nudge tries to alter people's behaviour in a predictable way without forbidding any options or significantly changing economic incentives. A nudge is not a legal requirement. Neither is it an economic sanction. Fines, taxes and subsidies are not nudges.

When used as a part of government policy, nudges must be open and transparent to the general public. Governments should be honest with the public and ensure that they explain why they have introduced a nudge, but still allow individuals to make a choice.

Nudges versus shoves

'Nudge' policies seek to lead people by providing them with helpful information and language that then allows them to make an informed choice. By contrast, 'shove' policies instruct people to behave in certain ways, often by responding to financial incentives and disincentives that reward or punish different decisions (see Table 2.4).

Government policies based on traditional economic theories have generally sought to shove people into altering their behaviour rather than to nudge them in the desired direction.

TEST YOURSELF 2.10

To what extent are fines and penalty points the most effective way of reducing the speed of motorists on the roads?

SYNOPTIC LINK

When studying the conditions of demand and supply in Chapter 3, and the problem of merit and demerit goods in Chapter 8, consider how 'nudge' policies could be more effective than the traditional 'shove' policies of taxation and subsidy.

Table 2.4 Nudges versus shoves

Nudge	Shove
• Provides information for people to respond to. • Creates positive social norms. • Opt-out schemes rather than opt-in schemes and default choices. • Active choosing by individuals.	• Uses taxation and subsidies to alter incentives and on occasion, in the case of taxes, to punish people. • Uses fines, laws banning activities and regulations.

APPLICATION OF ECONOMICS IN THE REAL WORLD 2.4

This case study has been extracted from a paper published by the UK cabinet's Behaviour Insights Team (BIT), published in 2013.

Applying behavioural insights to charitable giving

This paper explores new and innovative ways of increasing charitable giving. It recognises the important indirect benefits of charitable giving that recent behavioural research has begun to explore. This research shows that giving both time and money has large benefits for the well-being of the giver as well as the receiver.

Experiments have shown, for example, that individuals are happier when given the opportunity to spend money on others. Similarly, volunteering is associated with increased life satisfaction — not only among volunteers, but also in the wider community. Charitable giving is good for donors, for beneficiaries, and for society at large.

Four behavioural insights

Insight 1 is to **'make it easy'**. One of the best ways of encouraging people to give is to make it easy for people to do so. Making it easy can include:

- giving people the option to increase their future payments to prevent donations being eroded by inflation
- setting defaults that automatically enrol new senior staff into giving schemes (with a clear option to decline)
- using prompted choice to encourage people to become charitable donors.

Insight 2 is to **'attract attention'**. Making charitable giving more attractive to an individual can be a powerful way of increasing donations. This can include:

- attracting individuals' attention, for example by using personalised messages
- rewarding the behaviour you seek to encourage, for example through matched funding schemes
- encouraging reciprocity with small gifts.

Insight 3 is to **'focus on the social'**. We are all influenced by the actions of those around us, which means we are more likely to give to charity if we see it as the 'social norm'. Focusing on the social involves thinking about:

- using prominent individuals to send out strong social signals
- drawing on peer effects, by making acts of giving more visible to others within one's social group
- establishing group norms around which subsequent donors 'anchor' their own gifts.

Insight 4 is that **'timing matters'**. If you get your timing right, it can really help to increase charitable donations. This might include:

- ensuring that charitable appeals are made at the moments when they are likely to be most effective — for example, people are more likely to make a donation in December than January
- understanding that people may be more willing to commit to future (increases in) donations than equivalent sums today.

Follow-up questions

1 The BIT paper recommends 'establishing group norms around which subsequent donors "anchor" their own gifts'. Explain the two terms 'group norms' and 'anchor'.
2 Identify, from within the extract, two examples of nudge theory being applied.
3 Why is December a better month than January to ask for charitable donations?
4 To what extent can a positive social norm affect the willingness of an individual making a charitable donation?

SECTION 2.4 SUMMARY

- Nudge theory, choice architecture and framing lie at the heart of the ways in which behavioural architecture can influence economic policy making.
- Well-designed choice architecture can guide individuals towards making better decisions.
- Governments can improve social welfare by designing government programmes that select a default choice which is seen to be in an individual's best long-term interest.
- Decision making can be influenced when framed in a context.
- Nudge policies seek to lead individuals into making decisions whereas shove policies instruct using the power of the law.

Questions

1 Explain the hypothesis (or law) of diminishing marginal utility.

2 How does traditional economic theory seek to influence individual decision making?

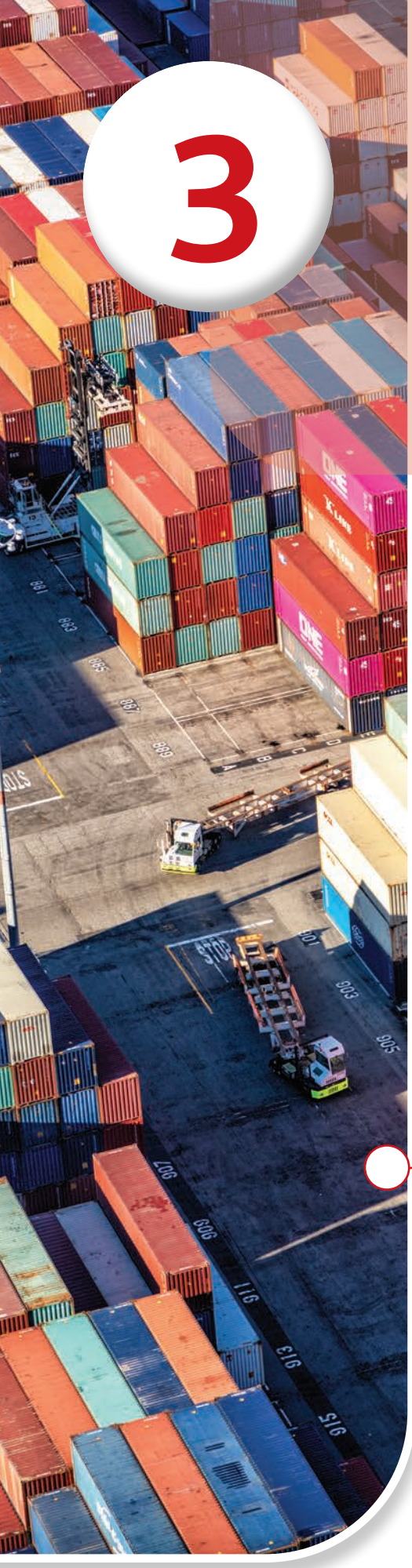
3 Explain how, in traditional economic theory, a rational individual can make a 'wrong' decision.

4 Explain why rational individuals may make satisficing rather than maximising choices.

5 Is it irrational to give money to charity? Discuss.

6 Explain, using examples, how government officials can use framing techniques to make policies more effective.

7 How can governments intervene in markets to help individuals make better choices?



3 Price determination in a competitive market

Chapter 1 introduced you to one of the fundamental economic problems: how to allocate scarce resources between competing uses in conditions in which there are limited resources and unlimited wants. In a market economy, resource allocation is undertaken by the price mechanism operating in the system of markets that make up the economy. This is also true in the 'market sector' of a mixed economy. However, in a mixed economy there is also a 'non-market sector' in which goods and services such as roads and police are produced and delivered to final users 'outside the market'. This chapter focuses on competitive markets in which the price mechanism operates. Chapter 5 examines the less competitive markets of monopoly and imperfect competition, as well as perfectly competitive market structures. The final chapter of Part 1 of this book, Chapter 8, introduces the various market failures in which the price mechanism either does not work at all or produces resource misallocation.

LEARNING OBJECTIVES

These are to understand:

- the nature of demand and supply in a competitive market
- the difference between a movement along a demand or a supply curve and a shift of a demand or a supply curve
- the concept of elasticity and the different elasticities you need to know
- how demand and supply curves are brought together in a supply and demand diagram
- market equilibrium and disequilibrium in a supply and demand diagram
- interrelationships between markets.

3.1 The determinants of demand

Demand and markets

Before we explain what determines the level of demand, we will first examine the nature of a **market** in which the **demand** for, and the **supply** of, goods and services interact with each other.

KEY TERMS

market a voluntary meeting of buyers and sellers with exchange taking place.

demand the quantity of a good or service that consumers are willing and able to buy at given prices in a given period of time.

supply the quantity of a good or service that producers are willing and able to sell at given prices in a given period of time.

A market is a voluntary meeting of buyers and sellers in which exchange takes place. Both buyer and seller have to be willing partners to the exchange. Markets do not have to exist in a particular geographical location. Whenever a good or service is voluntarily bought and sold, a market transaction occurs. In the past, market transactions shifted away from open-air street markets to take place in shops. Shops have higher overhead costs, but they offer a permanent site of exchange and a continuing relationship between sellers and buyers. In recent years, the growth of the internet has allowed 24/7 ecommerce. As a result, many markets, especially those in commodities, raw materials and financial services, have become truly global.

SYNOPTIC LINK

This chapter mainly focuses on product markets — the markets for the goods and services which consumers buy and businesses sell. However, there are a variety of other forms of market. For example, Chapter 6 looks at the market for the factor of production known as labour, while Chapter 15 considers the market for exchanging foreign currencies.

KEY TERMS

competitive markets markets in which the large number of buyers and sellers possess good market information and can easily enter or leave the market.

ruling market price (also known as **equilibrium price**) the price at which planned demand equals planned supply.

Competitive markets occur when there are a large number of buyers and sellers all passively accepting the **ruling market price**, which is set, not by individual decisions, but by the interaction of all those taking part in the market. Highly competitive markets lack entry and exit barriers. These barriers, which exist in non-competitive markets, include the costs of setting up, the costs of acquiring information about the market and the difficulties in gaining consumer recognition when there are established brands. The absence of these barriers means that new buyers and sellers can easily enter the market without incurring costs. In the same way, buyers and sellers can leave the market if they wish to. Competitive markets also exhibit a high degree of transparency — buyers and sellers can quickly find out what everyone else in the market is doing.

The relationship between price and quantity demanded

KEY TERM

effective demand the desire for a good or service backed by an ability to pay.

Households and firms operate simultaneously in two sets of markets. The first of these contains the goods markets, in which members of households demand and buy consumer goods and services produced and supplied by firms. But for household demand in the goods market to be an **effective demand** — that is, demand backed up by an ability to pay — households must first sell their labour, or possibly the services of any capital or land they own, in the markets for factors of production. These were briefly mentioned in Chapter 1. Households' roles are therefore reversed in goods markets and factor markets. In this chapter, we ignore factor markets and focus solely on the determinants of demand for consumer goods and services.

What a demand curve shows

A demand curve, such as the one illustrated in Figure 3.1, shows the relationship between the price of a good or service and the quantity of the good or service demanded at different prices. If the price starts off high, for

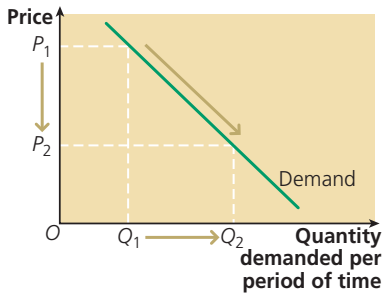


Figure 3.1 A market demand curve

example at P_1, household demand is Q_1. But if the price falls to P_2, quantity demanded increases to Q_2.

Demand for a good varies according to the time period being considered. For example, weekly demand is different from daily, monthly and annual demand. For this reason, the horizontal axis in Figure 3.1 is labelled 'Quantity demanded per period of time'. It is normal practice to use the label 'Quantity' on the horizontal axis of a demand curve diagram, as we do in the rest of this book, but this is an abbreviation. It always refers to a period of time.

TEST YOURSELF 3.1

A woman with an income of £15,000 a year dreams of owning a brand new Bugatti car priced at over £1 million. Explain why this is not an example of effective demand.

SYNOPTIC LINK

A microeconomic demand curve, such as the one illustrated in Figure 3.1, looks very similar to a macroeconomic aggregate demand curve, which is explained in Chapter 10. It is vital that you don't confuse the two. A demand curve represents an individual's demand for a single product, or the total demand for a single product, whereas an aggregate demand curve shows the planned expenditure in all markets in an economy. Likewise, don't confuse demand with consumption, which is a component of aggregate demand, also explained in Chapter 10.

Market demand and individual demand

Normally when economists refer to demand, they mean **market demand**. This is the quantity of a good or service that *all* the consumers in the market wish to, and are able to, buy at different prices. By contrast, **individual demand** is the quantity that a particular individual, such as yourself, would like to buy. The relationship between market and individual demand is simple. Market demand is just the sum of the demand of all the consumers in the market.

KEY TERMS

market demand the quantity of a good or service that all the consumers in a market are willing and able to buy at different market prices.

individual demand the quantity of a good or service that a particular consumer or individual is willing and able to buy at different market prices.

QUANTITATIVE SKILLS 3.1

Worked example: performing a percentage calculation

1 Calculate to 1 decimal place:
- **(a) 14% of £605**
- **(b) 4% of £4 million**
- **(c) 0.035% of £800**

For question (a), multiply £605 by 0.14, which gives the answer £84.70. Similarly, for question (b), multiply £4 million by 0.04, for which the answer is £160,000, and for question (c) multiply £800 by 0.00035, which gives the answer £0.28.

2 Express to 2 decimal places:
- **(a) 35 as a percentage of 450**
- **(b) 17 as a percentage of 130**
- **(c) £6 million as a percentage of £1 billion**

For question (a), divide 35 by 450 and convert into a percentage by multiplying by 100. This gives 7.777 recurring, which to 2 decimal places is 7.78%. Repeat the process for questions (b) and (c), which gives the answers 13.08% and 0.60%.

Shifts of a demand curve

Students often confuse a movement *along* a demand curve and a *shift* of a demand curve. A *movement along a demand curve* takes place only when the good's price changes. Provided the demand curve slopes downwards, a *fall* in price results in *more* of the good being demanded. This is sometimes called an extension of demand. Likewise, a contraction of demand occurs when a *rise* in price leads to *less* being demanded.

When we draw a market demand curve to show how much of the good or service households plan to demand at various possible prices, we assume that all the other variables that may also influence demand are held unchanged or constant. This is the *ceteris paribus* assumption, which means 'other things being equal'. Among the variables whose values are held constant or unchanged when we draw a demand curve are disposable income and tastes or fashion. Collectively, the variables (other than the good's own price) whose values determine planned demand are often called the **conditions of demand**. A change in a condition of demand shifts the demand curve to a new position.

KEY TERMS

condition of demand a determinant of demand, other than the good's own price, that fixes the position of the demand curve.

substitute goods alternative goods that could be used for the same purpose.

complementary goods when two goods are complements, they experience joint demand.

increase in demand a rightward shift of the demand curve.

decrease in demand a leftward shift of the demand curve.

STUDY TIP

You must understand the difference between a movement along a demand or supply curve and a shift of the curve. A movement along either curve implies a change in price alone and leads to an increase or decrease in quantity demanded or supplied. An increase or decrease in demand or supply results from a change in a factor other than price and leads to a change in the position of the curve.

The conditions of demand

The main conditions of demand are:

- the prices of **substitute goods** (or goods in competing demand) (see pages 56 and 70)
- the prices of **complementary goods** (or goods in joint demand) (see page 70)
- personal income (or, more strictly, personal disposable income, after tax and receipt of benefits)
- tastes and preferences
- population size, which influences total market size.

If any of the conditions of demand change, the position of the demand curve changes, shifting either rightwards or leftwards. Figure 3.2 illustrates a rightward shift of the demand curve, which is also called an **increase in demand**, showing that more of the good is demanded at all prices. For example, at a price of P_1, the quantity demanded increases from Q_1 to Q_2. Conversely, a leftward shift of demand (known as a **decrease in demand**) causes the quantity demanded to fall at all prices.

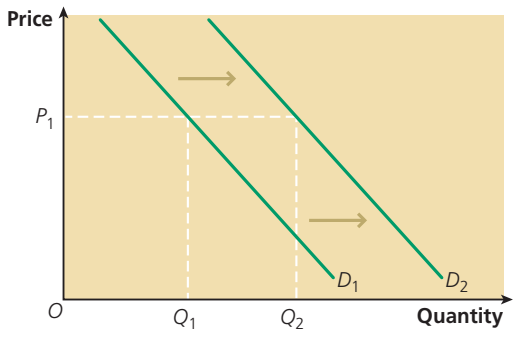


Figure 3.2 The effect of a rightward shift of demand

Events that might cause a rightward shift of a demand curve include:

- an increase in the price of a substitute good or a good in competing demand (see page 69 on the interrelationship between markets)
- a fall in the price of a complementary good or good in joint demand

- an increase in personal disposable income (but see the explanation of normal goods and inferior goods that follows)
- a successful advertising campaign making people think more favourably about the good
- an increase in population size.

TEST YOURSELF 3.2

Household income increases by 3% in a particular year. This causes demand for summer holidays to increase by 5%. Is this an example of an increase in demand or of a movement along a demand curve?

Normal goods and inferior goods

KEY TERMS

normal good a good for which demand increases as income rises and demand decreases as income falls.

inferior good a good for which demand decreases as income rises and demand increases as income falls.

When disposable income increases, demand for a good increases if the good is a **normal good**. However, some goods are examples of an **inferior good**, for which demand decreases as income increases.

To take an example, private car transport and bus travel are not just substitutes for each other. As people's incomes rise, demand for cars generally increases, while, at the same time, demand for bus travel usually falls. If people respond in this way to changes in income then private transport is a normal good, but certain forms of public transport are inferior goods. For an individual, whether a good is normal or inferior depends on personal income, tastes and, possibly, age. For young children, junk food such as sweets is usually a normal good, but later in life, tastes change and sweets may become an inferior good.

If a good is a normal good, an increase in income shifts the good's demand curve to the right. However, if the good is inferior for most people, its demand curve shifts to the left when income increases.



Junk food may be a normal good for a child and an inferior good for an adult

EXTENSION MATERIAL

Must demand curves always slope downwards?

Demand curves don't have to slope downwards, though they usually do. However, there are circumstances in which a demand curve may be horizontal or vertical, or indeed slope upwards, showing that more is demanded as the good's price increases.

There are a number of possible explanations for upward-sloping demand curves. Some of these are as follows:

- **Speculative demand.** If the price of a good such as housing, shares or a foreign currency starts to rise, people may speculate that in the near future the price will rise even further. In this situation, demand is likely to increase. In the case of house prices, young people who wish to become first-time buyers may scramble to buy houses even when prices are rising, fearing that if they wait, they may never be able to afford to get on the 'housing ladder'.
- **Goods for which consumers use price as an indicator of quality.** Consumers may lack accurate

information about the quality of some goods they want to buy, such as second-hand cars and computers. In this situation, a potential buyer may demand more as a good's price rises, believing that a high price means high quality.

- **Veblen goods.** Some companies try to sell their goods based on the fact that they cost more than those of their competitors. Veblen goods, named after the Norwegian-American economist Thorstein Veblen, are goods of exclusive or ostentatious consumption, or 'snob' goods. They are sometimes called positional goods, though strictly speaking, a positional good is so scarce that few people can ever acquire it. Some people wish to consume Veblen goods, such as Ferrari cars, as a signal of their wealth. The 'reassuringly expensive' advertising campaign for Stella Artois beer is another good example. Until around 2007 Interbrew, the Belgian company (now part of AB InBev) that then owned the Stella brand, decided to sell its beer as a premium brand. Interbrew hoped that high prices would attract more customers. However, if you look at the prices of Stella beer today, you will find that AB InBev has now changed tack, selling its beer on a 'stack 'em high, sell 'em fast' principle, at discounted prices.

SECTION 3.1 SUMMARY

- Demand means effective demand, based on ability as well as willingness to pay.
- A market demand curve shows how much of a good all the consumers in the market intend to buy at different prices.
- For most goods, demand curves slope downwards.
- The conditions of demand fix the position of the demand curve.
- If any of the conditions of demand change, the demand curve shifts to a new position.
- Movements along a demand curve must not be confused with a shift in the position of the curve.

3.2 Price, income and cross elasticities of demand

The meaning of elasticity

KEY TERM
elasticity the proportionate responsiveness of a second variable to an initial change in the first variable.

Whenever a change in one variable (such as a good's price) causes a change to occur in a second variable (such as the quantity of the good that households are prepared to demand), an **elasticity** can be calculated. The elasticity measures the proportionate responsiveness of the second variable to the change in the first variable. For example, if a 5% increase in price were to cause households to reduce their demand by more than 5%, demand would be elastic. In this example, a change in price induces a more than proportionate response by consumers. But if the response were less than a reduction of 5%, demand would be inelastic. And if the change in price were to induce exactly the same proportionate change in demand, demand would be neither elastic nor inelastic — this is called unit elasticity of demand.

Elasticity is a useful descriptive statistic of the relationship between two variables because it is independent of the units, such as price and quantity units, in which the variables are measured.

Although, in principle, economists could calculate a great many elasticities for all the economic relationships in which they are interested, the three demand elasticities you must know are:

- price elasticity of demand
- income elasticity of demand
- cross elasticity of demand.

The following formulae are used for calculating these elasticities:

$$\text{price elasticity of demand} = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in price}}$$

$$\text{income elasticity of demand} = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in income}}$$

$$\text{cross-elasticity of demand} = \frac{\text{percentage change in quantity of A demanded}}{\text{percentage change in price of B}}$$

STUDY TIP

Remember that elasticities are calculated by dividing the percentage change in quantity demanded (or supplied) by the percentage change in the variable that caused the change.

Price elasticity of demand

KEY TERM

price elasticity of demand measures the extent to which the demand for a good changes in response to a change in the price of that good.

Price elasticity of demand measures consumers' responsiveness to a change in a good's price. (It is sometimes called 'own price' elasticity of demand to distinguish it from cross-elasticity of demand for good A with respect to the price of B, which measures the responsiveness of demand for a particular good to a change in the price of a completely different good.)

STUDY TIP

You should apply elasticity analysis when assessing the effects of a shift of a demand or supply curve. The extent to which the good's price or equilibrium level of output changes depends on the price elasticity of the curve that has not shifted. For example, when the supply curve shifts leftwards, the price elasticity of the demand curve determines the extent to which the good's price and quantity change.

TEST YOURSELF 3.3

When the price of a small car is £15,000, there are 100,000 people in the UK who wish to buy it. When the price falls to £10,000, the number wanting to buy the car rises to 200,000. What does this information tell you about the market for small cars in the UK?

Infinite and zero price elasticity of demand

Horizontal and vertical demand curves have constant elasticities at all points on the curve. A horizontal demand curve, such as the demand curve in Figure 3.3(a), is infinitely elastic or perfectly elastic. At the other extreme,

the vertical demand curve in Figure 3.3(b) is completely inelastic or perfectly inelastic, displaying a zero price elasticity of demand at all points on the curve. When the price falls, for example from P_1 to P_2, the quantity demanded is unchanged.

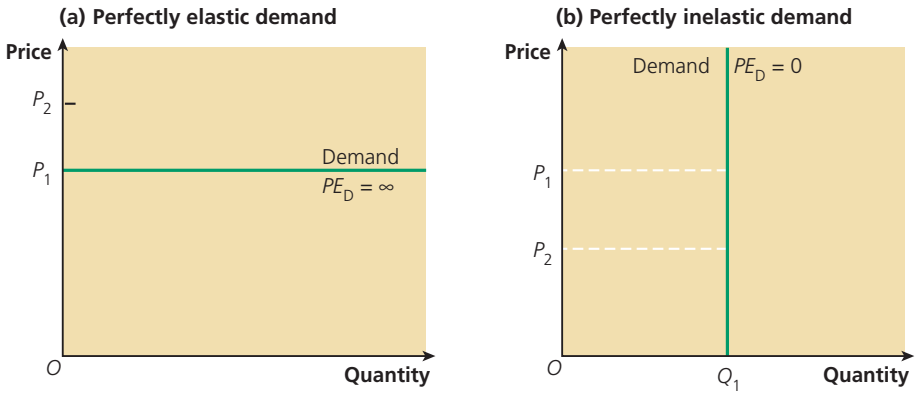


Figure 3.3 Horizontal and vertical demand curves

Figure 3.4 summarises the five demand curves you need to know.

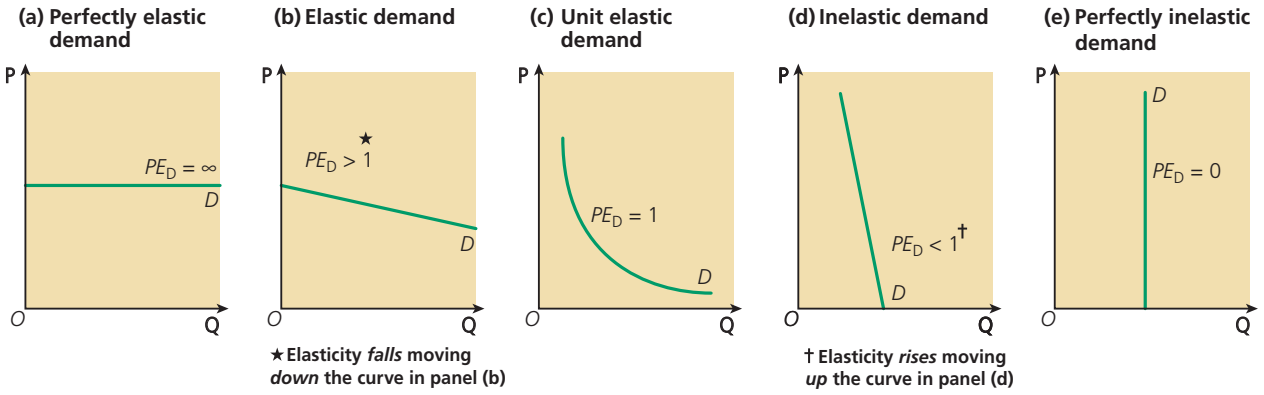


Figure 3.4 Five demand curves you need to know

Factors determining price elasticity of demand

Substitutability

Substitutability is the most important determinant of price elasticity of demand. When a substitute exists for a product, consumers respond to a price rise by switching expenditure away from the good and buying a substitute whose price has not risen. When very close substitutes are available, demand for the product is highly elastic. Conversely, demand is likely to be inelastic when no substitutes or only poor substitutes are available.

Percentage of income

The demand curves for goods or services on which households spend a large proportion of their income tend to be more elastic than those of small items that account for only a small fraction of income. This is because, for items on which only a very small fraction of income is spent, particularly for those which are rarely purchased, people hardly notice the effect of a change in price on their income. The same is not true for 'big-ticket' items such as a new car or an overseas holiday.

Necessities or luxuries

It is sometimes said that the demand for necessities is price inelastic, whereas the demand for luxuries is elastic. This statement should be treated with caution. When no obvious substitute exists, demand for a luxury good may be inelastic, while at the other extreme, demand for particular types of basic foodstuff is likely to be elastic if other staple foods are available as substitutes. It is the existence of substitutes that really determines price elasticity of demand, not the issue of whether the good is a luxury or a necessity.

The 'width' of the market definition

The wider the definition of the market under consideration, the lower the price elasticity of demand. Thus the demand for the bread produced by a particular bakery is likely to be more elastic than the demand for bread produced by all bakeries. This is because the bread baked in other bakeries provides a number of close substitutes for the bread produced in just one bakery. And if we widen the possible market still further, the elasticity of demand for bread produced by all the bakeries will be greater than that for food as a whole.

Time

The time period in question will also affect the price elasticity of demand. For many goods and services, demand is more elastic in the **long run** than in the **short run** because it takes time to respond to a price change. For example, if the price of an electric-powered car falls relative to the price of a petrol-engine car, it will take time for motorists to respond because they will be 'locked in' to their existing investment in petrol-engine cars.

In other circumstances, the response might be greater in the short run than in the long run. A sudden rise in the price of petrol might cause motorists to economise in its use for a few weeks before getting used to the price and drifting back to their old motoring habits.

KEY TERMS

short run the time period in which at least one factor of production is fixed and cannot be varied.

long run the time period in which no factors of production are fixed and all the factors of production can be varied.

APPLICATION OF ECONOMICS IN THE REAL WORLD 3.1

Elasticity and tobacco taxation

Various studies have calculated the price elasticity of demand for cigarettes of different groups in society, such as the young and the old, and men and women.

A World Bank review concluded that price rises of about 10% would on average reduce tobacco consumption by about 4% in richer countries. Smokers in poorer nations tend to be more sensitive to price changes.

Reviewing 86 studies, Gallet and List found a mean price elasticity of –0.48, meaning that, on average, a 10% increase in price will be followed by a decrease in consumption of 4.8%. They also found greater responsiveness among younger people, with an average price elasticity of –1.43 for teenagers, –0.76 for young adults and –0.32 for adults. They found an average price sensitivity of –0.50 for men and –0.34 for women. Studies have also tended to show greater price sensitivity among low-income groups.

Follow-up questions

1. What is the formula which is used for calculating price elasticity of demand?
2. Name the other two types of elasticity of demand besides price elasticity of demand.
3. Suggest two reasons why adult smokers may be less responsive to a rise in the price of cigarettes than teenage smokers.
4. Most of the elasticity statistics quoted above lie between zero and –1. Discuss the significance of this for governments.

QUANTITATIVE SKILLS 3.2

Worked example: performing an elasticity calculation

People's average incomes fall from £1,000 a week to £600 a week. As a result, demand for potatoes increases from 1 million tonnes to 1.2 million tonnes a week. Calculate the income elasticity of demand for potatoes.

The formula for calculating income elasticity of demand is:

$$\text{income elasticity of demand} = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in income}}$$

The percentage change in quantity demanded is +20%. The percentage change in income is –40%. Placing these figures into the formula:

$$\text{income elasticity of demand} = \frac{+20\%}{-40\%} = -0.5$$

The minus sign indicates that the good is an inferior good. The number 0.5 indicates that demand is inelastic.

Price elasticity of demand, total consumer expenditure and firms' total revenue

As an alternative to using the formula to calculate price elasticity of demand between two points on a demand curve, a simple rule can be used to determine the general nature of the elasticity between the two points:

- If total consumer expenditure increases in response to a price fall, demand is elastic.
- If total consumer expenditure decreases in response to a price fall, demand is inelastic.
- If total consumer expenditure remains constant in response to a price fall, demand is neither elastic nor inelastic, i.e. elasticity = unity (or since the demand curve slopes downwards, the elasticity is minus unity or –1).

Consider, for example, Figure 3.5, which shows an elastic demand curve D. At price P_1, total consumer expenditure is shown by the rectangle bounded by P_1, a, Q_1 and O. When the price falls to P_2, the consumer expenditure rectangle changes to the area bounded by P_2, b, Q_2 and O. Clearly, the second of these rectangles is larger than the first rectangle, so total consumer expenditure increases, following a fall in price, when the demand curve is elastic.

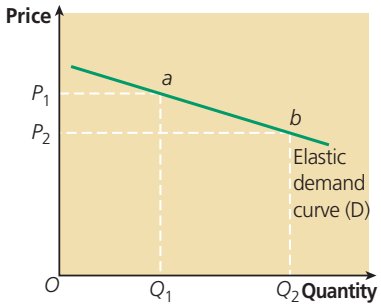


Figure 3.5 The effect of a price fall on total consumer expenditure when demand is elastic

STUDY TIP

Total consumer expenditure is exactly the same as firms' total sales revenue, so if we prefer, we can state the rule in terms of revenue rather than consumer expenditure.

EXTENSION MATERIAL

The slope of a demand curve and its elasticity

Although economists typically illustrate price elastic demand with a shallow-sloping demand curve and price inelastic demand with a steep one, it is interesting to note that the elasticity of demand varies along the length of a demand curve. Take a careful look at the two demand curves in Figure 3.6. In Figure 3.6(a), a straight line (or linear) demand curve has been drawn. Obviously, a straight line has a constant slope. But although the slope is the same at all points on the curve, the elasticity is not.

Moving along a linear downward-sloping demand curve, the price elasticity of demand falls from point to point along the curve. Demand is elastic (or greater than unity) at all points along the top half of the curve. Elasticity equals unity exactly half way along the curve, falling below unity and towards zero along the bottom half of the curve.

If elasticity falls from point to point moving down a linear demand curve, it follows that a non-linear curve (i.e. a curved line) is needed to show the same elasticity at all points on the curve. Figure 3.6(b) shows a demand curve with a constant elasticity of 1 at all points on the curve: that is, elasticity equals unity at all points on the curve. Mathematicians call this a rectangular hyperbola. Whenever the price falls, the proportionate change in quantity demanded exactly equals the proportionate change in price. In this case, consumer expenditure remains unchanged following a rise or fall in price.

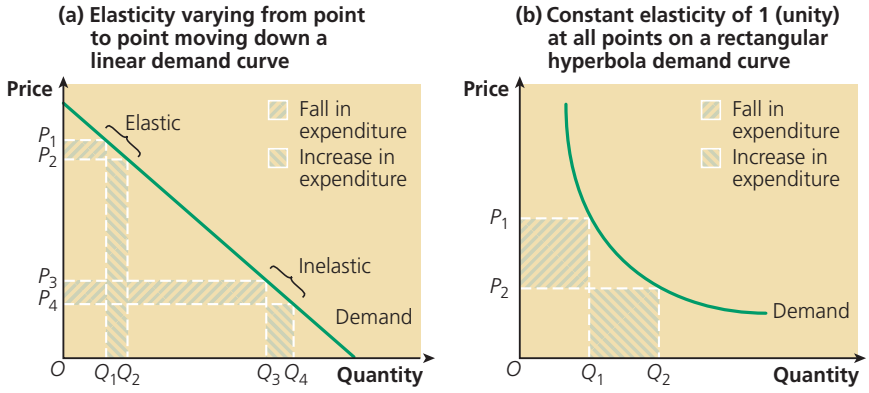


Figure 3.6 Price elasticity of demand and the slope of the demand curve

TEST YOURSELF 3.4

Draw a diagram to illustrate what happens to total consumer expenditure in the event of a price fall when the demand curve is inelastic.

Income elasticity of demand

KEY TERM

income elasticity of demand measures the extent to which the demand for a good changes in response to a change in income; it is calculated by dividing the percentage change in quantity demanded by the percentage change in income.

The nature of **income elasticity of demand** — which measures how demand responds to a change in income — depends on whether the good is a normal good or an inferior good.

When disposable income increases, a demand curve shifts rightwards, but only if the good is a normal good, for which demand increases as income increases. However, some goods are inferior goods, for which demand decreases as income increases, and an increase in income shifts the demand curve leftwards.

Income elasticity of demand is *always* negative for an inferior good and positive for a normal good. This is because the quantity demanded of an inferior good falls as income rises, whereas the quantity demanded of a normal good rises with income.

Normal goods can be further divided into superior goods or luxuries, for which the income elasticity of demand is greater than +1, and necessities, with an income elasticity lying between zero and +1. Although the quantity demanded of a normal good always rises with income, it rises by a greater percentage for a superior good (such as a luxury car). Conversely, demand for a basic good or necessity such as shoe polish rises by a smaller percentage than income.

TEST YOURSELF 3.5

The UK's income elasticity for overseas holidays is +1.6. What does this tell you about UK demand for overseas holidays?

The size and sign (positive or negative) of income elasticity of demand affect how a good's demand curve shifts following a change in income.

Cross elasticity of demand

KEY TERM

cross elasticity of demand measures the extent to which the demand for a good changes in response to a change in the price of another good; it is calculated by dividing the percentage change in quantity demanded by the percentage change in the price of another good.

Cross elasticity of demand (or cross-price elasticity of demand) measures how the demand for one good responds to changes in the price of another good. The cross elasticity of demand between two goods or services indicates the nature of the demand relationship between the goods. There are three possibilities:

- complementary goods (or joint demand)
- substitutes (or competing demand)
- an absence of any discernible demand relationship.

Cars and petrol or diesel fuel, for example, are complementary goods: they are in joint demand. A significant increase in fuel prices, such as the 30.7% increase in petrol prices in the year to March 2022, will have some effect on the demand for cars, though the effect may not be great. By contrast, private car travel and bus travel are substitute goods. A significant increase in the cost of running a car will cause some motorists to switch to public transport — provided its price does not rise by a similar amount as well.

As with the case of income elasticity of demand, the size and sign (positive or negative) of cross elasticity of demand affect how a good's demand curve shifts following, in this case, a change in the price of another good. For example, a cross elasticity of demand of +0.3 for bus travel with respect to the price of running a car indicates that a 10% increase in the cost of private motoring would cause the demand for bus travel to increase by just 3%. For most demand relationships between two goods, cross elasticities of demand are inelastic rather than elastic, both when the goods are in joint demand and when they are substitutes.

Complementary goods, or goods which are demanded together, such as bicycles and bike lamps, have negative cross elasticities of demand. A rise in the price of one good leads to a fall in demand for the other good. Suppose, for example, that the cross elasticity of demand for bike lamps with respect to the price of new bicycles is –0.5: this tells us that a 10% increase in the price of a new bicycle leads to a 5% fall in the demand for bike lamps.

By contrast, the cross elasticity of demand between two goods which are substitutes for each other is positive. A rise in the price of one good causes demand to switch to the substitute good whose price has not risen. Demand for the substitute good increases. For example, a new bicycle and a new motor scooter are substitutes for each other. If the cross elasticity of demand for a new

TEST YOURSELF 3.6
The price of a gaming console for a particular games provider rises by 30%. In subsequent years, the demand for games cartridges for this system falls by 10%. What does this tell you about the cross elasticity of demand between the two products?

bicycle with respect to the price of a motor scooter is +0.4, this tells us that a 10% increase in the price of a new motor scooter will lead to a 4% increase in the demand for new bicycles as consumers switch between the two types of private transport.

If we select two goods at random — for example, pencils and suitcases — the cross elasticity of demand between the two goods will be zero. When there is no discernible demand relationship between two goods, a rise in the price of one good will have no measurable effect upon the demand for the other. The cross elasticity of demand is zero, unless, of course, both items make up an important part of household expenditure.

STUDY TIP
Elasticity basically means responsiveness. Demand elasticities measure how consumers respond to a change in a good's price, income, or the price of another good. You should learn the formulae for each type of elasticity — for example, income elasticity of demand — and avoid making three basic mistakes when using them. The mistakes are:

- missing out the word 'percentage' (or the % sign) from the formula
- writing the formula 'upside down'
- confusing the different elasticities.

SECTION 3.2 SUMMARY

- Elasticity means responsiveness.
- There are three important demand elasticities: price, income and cross elasticity of demand.
- The slope of a demand curve is not the same as price elasticity of demand.
- It is important to understand the determinants of all the elasticities you need to know.
- It is important to learn all the elasticity formulas, and not confuse them.
- You should be able to interpret elasticity statistics.

3.3 The supply of goods and services

Market supply

KEY TERM
market supply the quantity of a good or service that all the firms in a market plan to sell at given prices in a given period of time.

Normally when economists refer to supply, they mean market supply. **Market supply** is the quantity of a good or service that all the firms or producers in the market plan to sell at different prices. By contrast, supply by a single firm is the quantity that a particular firm within the market would like to sell. As with demand, the relationship between the two is simple. Market supply is just the sum of the supply of all the firms or producers in the market at different market prices.

Figure 3.7 shows a market supply curve which indicates that as a good's price rises, more is supplied. If the price starts off low, for example at P_1, firms are willing to supply Q_1. But if the price rises to P_2, planned supply increases to Q_2.

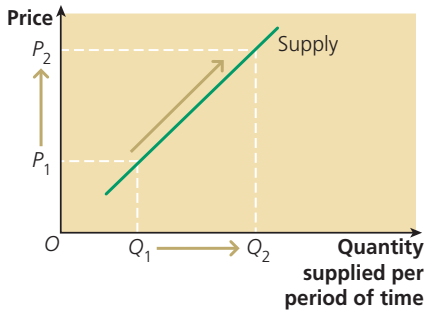


Figure 3.7 A market supply curve

The main reason for upward-sloping supply curves stems from the profit-maximising objective which economists assume firms have. If we assume that a firm always aims to make the biggest possible profit, it follows that a firm will only want to supply more of a good if it is profitable to do so.

For a firm, **profit** is the difference between the **total revenue** the firm receives when selling the goods or services it produces and the costs of producing the goods. Assuming firms do not change their size or scale, the cost of producing extra units of a good generally increases as firms produce more of the good. As a result, it is unprofitable to produce and sell extra units of a good unless the price rises to compensate for the extra cost of production. Rising prices will also encourage new firms to enter the market. The result is the upward-sloping market supply curve we have illustrated.

KEY TERMS

profit the difference between total sales revenue and total costs of production.

total revenue all the money received by a firm from selling its total output.

As with demand, the supply of a good varies according to the time period being considered. Hence the words 'Quantity supplied per period of time' on the horizontal axis in Figure 3.7. In later diagrams, this is shortened to 'Quantity'. But again, as with demand, remember that this is an abbreviation.

TEST YOURSELF 3.7

A farmer sells 100 sheep at a price of £20 per sheep. What is the farmer's total revenue and why is this not the same as the farmer's profit?

SYNOPTIC LINK

Microeconomic supply curves look very similar to aggregate supply curves, which are explained in Chapter 10. Don't confuse the two. Microeconomic supply curves show the willingness of individual firms or the total number of firms in one market to supply a single product, whereas aggregate supply curves show the total planned output of all firms in an economy taken together.

Shifts of a supply curve

Earlier in the chapter, we saw that a market demand curve shows how much all the consumers in the market plan to buy at different prices of the good, assuming all the other factors that influence demand remain constant. These 'other factors' were called the conditions of demand and we explained how, if any of them change, the demand curve shifts to a new position.

In exactly the same way, a market supply curve shows the quantities of the good that all the firms in the market plan to supply at different possible prices, assuming the **conditions of supply** remain unchanged. Again, if the *ceteris paribus* assumption no longer holds, one or more of the conditions of supply change and the supply curve shifts to a new position.

KEY TERM

condition of supply a determinant of supply, other than the good's own price, that fixes the position of the supply curve.

The conditions of supply

The main conditions of supply are:

- costs of production, including
  - wage costs
  - raw material costs

KEY TERMS

increase in supply a rightward shift of the supply curve.

decrease in supply a leftward shift of the supply curve.

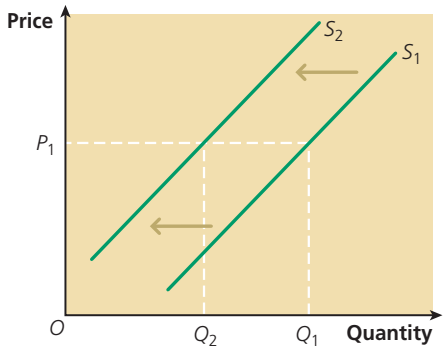


Figure 3.8 A leftward shift of the supply curve

- energy costs
- costs of borrowing
- technical progress
- taxes imposed on firms, such as VAT, excise duties and the business rate
- subsidies granted by the government to firms.

As we have noted, if any of the conditions of supply change, the supply curve shifts to a new position. As with demand, a rightward shift of supply is known as an **increase in supply**, whereas a leftward shift is known as a **decrease in supply**. An increase in wage costs, which for many firms are the most important cost of production, shifts the supply curve leftwards (or upwards). Firms reduce the quantity of the good they are prepared to supply because production costs have risen. For example, when the price is P_1 in Figure 3.8, a leftward shift of supply from S_1 to S_2 causes the quantity firms are prepared to supply to fall from Q_1 to Q_2. An expenditure tax such as VAT imposed by the government on firms would have a similar effect to an increase in costs of production.

Supply curves tend to shift rightwards when technical progress occurs, reducing production costs, or when firms enter the market. A subsidy given to firms by the government, being similar to a reduction in costs of production, would also shift the supply curve to the right.

TEST YOURSELF 3.8

Having set the price of bread at £3 a loaf, bread shops bake 10 million loaves, which they then try to sell. However, 8 million of these loaves remain unsold. What is likely to happen next in the bread market?

EXTENSION MATERIAL

Expenditure taxes and subsidies

A supply curve shifts leftwards (or upwards) when the government imposes an expenditure tax such as customs and excise duties or VAT on firms. From a firm's point of view, the tax is similar to a rise in production costs. Firms try to pass the tax on to consumers by increasing the price of the good. For this reason, expenditure taxes provide examples of indirect taxes. The higher price charged means consumers indirectly pay the tax, even though the firms and not the consumers pay the tax to the government.

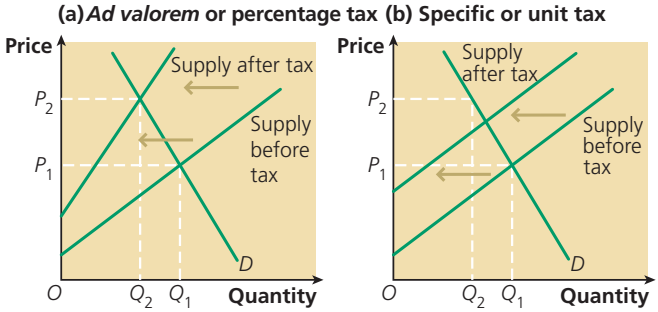


Figure 3.9 An expenditure tax shifting a supply curve

How the supply curve shifts depends on whether the tax that firms are forced to pay is an *ad valorem* tax or a specific tax. In the case of an *ad valorem* tax such as VAT, which is levied at the same percentage rate (e.g. 20%) on the price, the new supply curve is steeper than the old supply curve. This is shown in Figure 3.9(a). If a good is priced at £1, 20% of the price without the tax is 20p. However, if the price of a good is £2, the government collects 40p of tax revenue for each unit of the good sold.

But in the case of a specific tax or unit tax, such as the excise duty levied on tobacco, the tax levied does not depend on the good's price. Because of this, the new and old supply curves are parallel to each other, separated, as Figure 3.9(b) illustrates, by the size of the tax levied on each unit of the good. When an indirect tax is imposed on a good, the supply curve shifts vertically upwards by the amount of the tax.

A subsidy given by the government to producers has the opposite effect to an expenditure tax; it shifts the supply curve to the right. In the case of

a specific subsidy, which is illustrated in Figure 3.10, the sum of money paid to firms for each unit of the good produced is the same whatever the price of the good and hence the vertical distance between the two supply curves equals the subsidy per unit. By contrast, the size of the subsidy would vary if the subsidy were dependent on the price of the good.

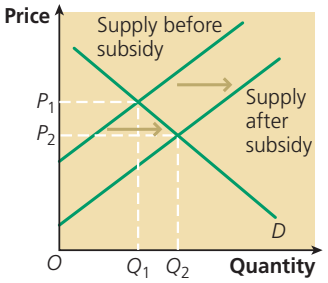


Figure 3.10 A specific or unit subsidy shifting a supply curve

TEST YOURSELF 3.9

Suppose the government subsidises private education by giving parents £5,000 a year if their child attends a private school. Would this shift the demand curve for, or the supply curve of, private education? Justify your answer.

SECTION 3.3 SUMMARY

- A market supply curve shows how much of a good all the firms in the market intend to supply at different prices.
- Supply curves usually slope upwards because higher prices lead to higher profits, encouraging existing firms to produce more and attracting new firms into the market.
- The conditions of supply fix the position of the supply curve.
- If any of the conditions of supply change, the supply curve shifts to a new position.
- Movements along a supply curve must not be confused with a shift in the position of the curve.

3.4 Price elasticity of supply

KEY TERM
price elasticity of supply measures the extent to which the supply of a good changes in response to a change in the price of that good.

In contrast to demand elasticities explained earlier in the chapter, there is only one supply elasticity you need to know. This is **price elasticity of supply**, which measures how the supply of a good responds to an initial change in a good's price.

The formula for price elasticity of supply is:

$$\text{price elasticity of supply} = \frac{\text{percentage change in quantity supplied}}{\text{percentage change in price}}$$

Just as with demand curves, you must not confuse the *slope* of a supply curve with its *elasticity*. Upward-sloping *straight-line* (linear) supply curves display the following price elasticities:

- If the supply curve intersects the price axis, the curve is elastic at all points, though elasticity falls towards unity as you move from point to point up the curve.
- If the supply curve intersects the quantity axis, the curve is inelastic at all points, though elasticity rises towards unity as you move from point to point up the curve.
- If the supply curve passes through the origin, elasticity equals unity (+1) at all points on the curve.

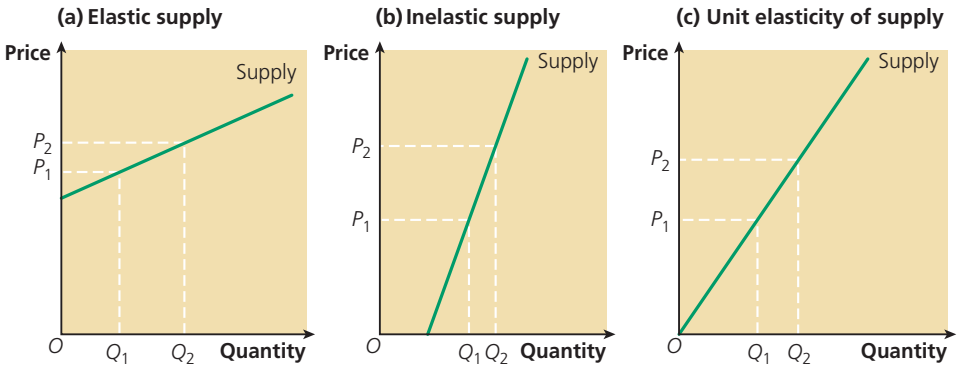


Figure 3.11 Price elasticity of supply and linear supply curves

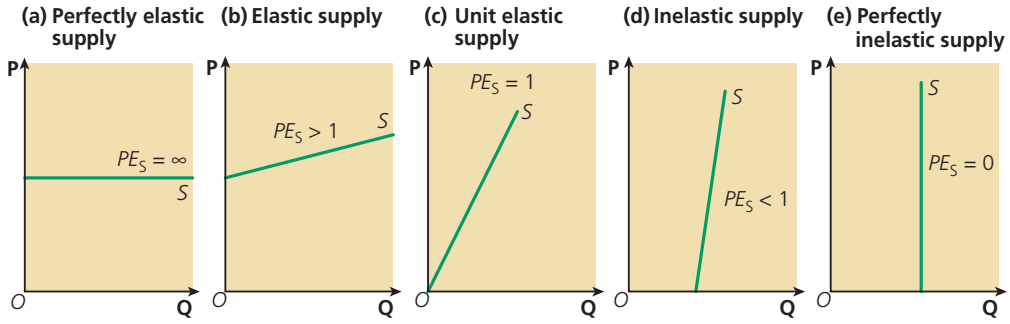


Figure 3.12 The five linear supply curves you should know

TEST YOURSELF 3.10

If the price of a good with a price elasticity of supply of 2.5 increases by 10%, the quantity supplied will:

A fall by 25%

B rise by 25%

C fall by 40%

D rise by 0.4%

Explain your answer.

STUDY TIP

You should understand why price elasticity of supply is usually positive and why price elasticity of demand is usually negative. This is because we assume that a standard demand curve slopes downwards left to right, showing that as the price of a product falls, the quantity demanded increases. Conversely, a standard supply curve slopes upwards left to right, showing that an increase in the price of a product usually leads to an increase in the quantity supplied.

The factors determining price elasticity of supply

The length of the production period

If firms can convert raw materials into finished goods very quickly (e.g. in just a few hours or days), supply will tend be more elastic than when several months are involved in production, as with many agricultural goods.



The supply of most agricultural goods is price inelastic

The availability of spare capacity

When a firm possesses spare capacity, and if labour and raw materials are readily available, production can generally be increased quickly in the short run.

The ease of accumulating stocks

When stocks of unsold finished goods are stored at low cost, firms can respond quickly to a sudden increase in demand. Alternatively, firms can respond to a price fall by diverting current production away from sales and into stock

accumulation. The ease with which stocks of raw materials or components can be bought from outside suppliers and then stored has a similar effect.

The ease of switching between alternative methods of production

When firms can quickly alter the way they produce goods — for example, by switching between the use of capital and labour — supply tends to be more elastic than when there is little or no choice. In a similar way, if firms produce a range of products and can switch raw materials, labour or machines from one type of production to another, the supply of any one product tends to be elastic.

The number of firms in the market and the ease of entering the market

Generally, the more firms there are in the market, and the greater the ease with which a firm can enter or leave, the greater the elasticity of supply.

Time

We have already noted that demand is more elastic in the long run than in the short run because it takes time to respond to a price change. The same is true for supply. Figure 3.13 shows three supply curves of increasing elasticity, S_1, S_2 and S_3, which illustrate respectively market period supply, short-run supply and long-run supply.

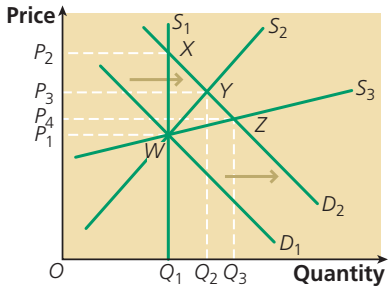


Figure 3.13 The effect of the time period upon price elasticity of supply

- **Market period supply.** The market period supply curve S_1 is shown by a vertical line. S_1 depicts the situation facing firms following a sudden and unexpected rightward shift of demand from D_1 to D_2. When surprised by a sudden increase in demand, firms cannot immediately increase output. In the market period, supply is completely inelastic, and the price rises from P_1 to P_2 to eliminate the excess demand brought about by the rightward shift of the demand curve.
- **Short-run supply.** The higher price means that higher profits can be made, creating the incentive for firms to increase output. In the short run, firms increase output by hiring more variable factors of production, such as labour. The short-run increase in output is shown by the movement up the short-run supply curve, S_2. The short-run supply curve is more elastic than the market period supply curve, S_1. In the short run, supply increases to Q_2, and the price falls from P_2 to P_3.
- **Long-run supply.** If firms believe the increase in demand will be long-lasting, and not just a temporary phenomenon, they may increase the scale of production by employing more capital and other factors of production that are fixed in the short run, but variable in the long run. When this happens, firms move along the long-run supply curve S_3. Output rises to Q_3 and the price falls once again, in this case to P_4.

STUDY TIP

You should understand why, for most goods, both the demand curve and the supply curve are more price elastic in the long run than in the short run.

In a competitive industry with low or non-existent barriers to entry, elasticity of supply is greater in the long run than in the short run, because in the long run firms can enter or leave the market. Short-run supply is less elastic because supply is restricted to the firms already in the industry.

APPLICATION OF ECONOMICS IN THE REAL WORLD 3.2

Housing market elasticities in the UK

UK households have an income elasticity of demand for housing that exceeds +1. However, demand for housing is price inelastic. These demand elasticities, combined with a low price elasticity of supply for housing, push the UK's housing market towards long-term rising prices.

To prevent rapid price rises, new housing would need to have a price elasticity of supply of +10 for supply to equal demand in the long term. But if the price elasticity of supply for new housing remains low, as Table 3.1 shows, house prices will never be stable in the UK when the demand for housing is increasing. Prices are also likely to be unstable when both demand and supply are highly price inelastic.

Table 3.1 Price elasticity of supply in the housing market for different countries

Country	Price elasticity of supply
Canada	+1.2
UK	+0.4
USA	+2.0
France	+0.3
Ireland	+0.6

Follow-up questions

1 State the formula used for measuring price elasticity of supply.
2 Distinguish between the slope and the elasticity of a supply curve.
3 Suggest why the price elasticity of supply of new houses is lower in the UK than in the USA.
4 'To prevent rapid price rises, new housing would need to have a price elasticity of supply of +10 for supply to equal demand in the long term.' Explain this statement.

EXTENSION MATERIAL

A closer look at perfectly elastic demand and supply

Figure 3.14 shows a perfectly elastic demand curve and a perfectly elastic supply curve. (These can also be labelled infinitely elastic demand and infinitely elastic supply.) Although the two parts of Figure 3.14 appear to be identical (apart from the labels), this is misleading. The apparent similarity disguises a significant difference between perfectly elastic demand and perfectly elastic supply. In Figure 3.14(a), demand is infinitely elastic at all prices on or *below* the demand curve, though if the price rises *above* the demand curve (e.g. from P_1 to P_2), the amount demanded immediately falls to zero. This is because perfect substitutes are available when demand is perfectly price elastic. Customers cease to buy the good as soon as the price rises *above* the demand curve, switching spending to the perfect substitutes whose prices have not changed.

By contrast, in Figure 3.14(b), supply is infinitely elastic at all prices on or *above* the supply curve, though if the price falls *below* the supply curve (e.g. from P_1 to P_2), the amount supplied immediately drops to zero. P_1 is the minimum price acceptable to firms. If they are paid this price (or any higher price), firms stay in the market. The incentive to stay in the market disappears at any lower price and firms leave the market, unable to make sufficient profit.

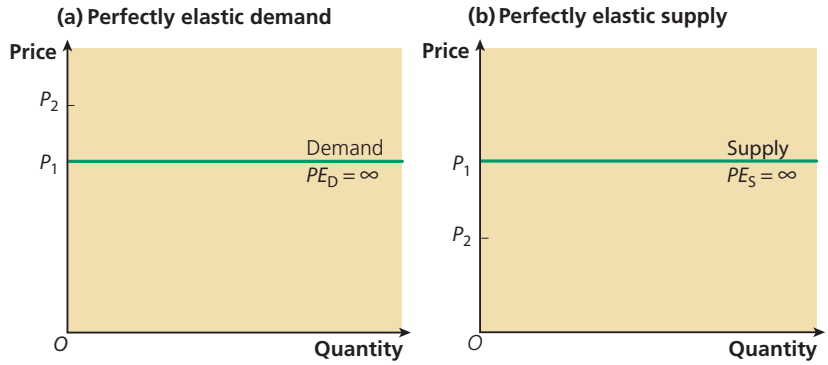


Figure 3.14 Perfectly price elastic demand and supply curves

TEST YOURSELF 3.11
Draw a diagram which shows a demand curve shifting rightwards along a supply curve with a zero price elasticity of supply. Explain what happens to the equilibrium price and quantity. Suggest a real-world market which the diagram illustrates.

SECTION 3.4 SUMMARY

- There is only one supply elasticity you need to know: price elasticity of supply.
- The slope of a supply curve is not the same as price elasticity of supply.
- It is important to understand the determinants of price elasticity of supply.
- You should be able to interpret price elasticity of supply statistics.

3.5 The determination of equilibrium market prices

The interaction of demand and supply

We now bring together the market demand and market supply curves explained earlier in the chapter to see how the equilibrium price is achieved in a competitive market within the economy. The market we will look at is the tomato market. Its essential features are shown in Figure 3.15.

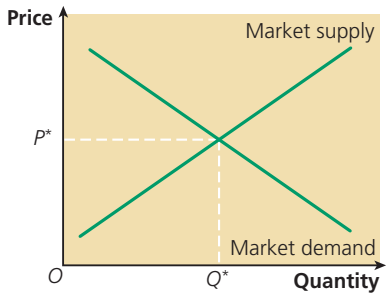


Figure 3.15 Market equilibrium in the tomato market

The market demand curve in Figure 3.15 shows how many tomatoes all the consumers in the market plan to purchase at different prices in a particular period of time. The market supply curve shows how many tomatoes all the farmers and firms in the market wish to supply at different prices in the same time period. The equilibrium market price, P^* in Figure 3.15, is located where the market demand curve for tomatoes intersects or cuts through the market supply curve of tomatoes. We will now look at the concept of **equilibrium** (and its opposite, **disequilibrium**) in greater depth.

KEY TERMS
equilibrium a state of rest or balance between opposing forces.
disequilibrium a situation in which opposing forces are out of balance.

Market equilibrium and disequilibrium

The concepts of equilibrium and disequilibrium are important in economic theory and analysis. You should think of equilibrium as a *state of rest* or a *state of balance between opposing forces* and of disequilibrium as a situation of the opposing forces being out of balance. In a market, the opposing forces are supply and demand. **Market equilibrium** in Figure 3.15 occurs where the demand curve and the supply curve cross each other. At price P^*, households *plan* to demand exactly the same quantity of tomatoes that firms *plan* to supply. P^* is therefore the equilibrium price, with Q^* being the equilibrium quantity.

In summary, **market disequilibrium** occurs when:

- planned demand < planned supply, in which case the price falls, or when
- planned demand > planned supply, in which case the price rises.

Market equilibrium occurs when:

- planned demand = planned supply, in which case the price does not change.

KEY TERMS
market equilibrium a market is in equilibrium when planned demand equals planned supply, where the demand curve crosses the supply curve.
market disequilibrium exists at any price other than the equilibrium price, when either planned demand < planned supply or planned demand > planned supply.

SYNOPTIC LINK

Refer back to the mention of 'ruling market price' on page 46. This tends to be the price at market equilibrium.

STUDY TIP

It is important to understand the concepts of equilibrium and disequilibrium in economics. These concepts help to explain why markets move from one state of rest to another. You will come across many other examples besides market equilibrium and disequilibrium. In your later studies, look out for profit-maximising equilibrium, equilibrium wage rate, equilibrium national income (or macroeconomic equilibrium) and balance of payments equilibrium.

TEST YOURSELF 3.12

What is meant by the equilibrium price of a good?

How excess demand and excess supply lead to changes in price

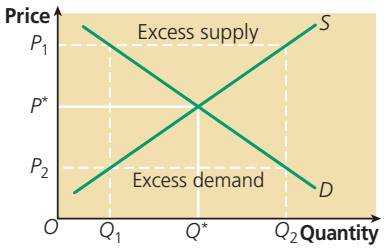


Figure 3.16 Disequilibrium and equilibrium in the tomato market

KEY TERMS

excess supply when firms wish to sell more than consumers wish to buy, with the price above the equilibrium price.

excess demand when consumers wish to buy more than firms wish to sell, with the price below the equilibrium price.

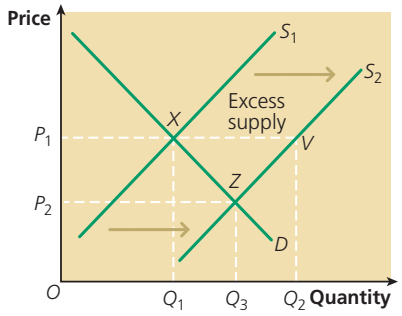


Figure 3.17 The effect of a rightward shift of the market supply curve of tomatoes

It is impossible at most prices for both households and firms simultaneously to fulfil their market plans. In Figure 3.16, P_1 is a disequilibrium price for tomatoes because the tomato growers and sellers cannot fulfil their plans at this price. When price is P_1 in Figure 3.16, firms would like to supply Q_2, but households are only willing to purchase Q_1.

To explain this further, it is useful to divide the market into two 'sides' — the short side and the long side. When the price is P_1, households, or the people wishing to buy tomatoes, are on the short side of the market, while tomato producers are on the long side. The economic agents on the short side can always fulfil their market plans, but those on the long side cannot. Thus, when the price is P_1, households can purchase exactly the quantity of tomatoes they wish to, namely Q_1. Farmers and other tomato producers, however, are in a different situation. They would like to sell Q_2, but can only sell Q_1, as long as the price remains at P_1. The difference between Q_2 and Q_1 is **excess supply** or unsold stock.

The market is also in disequilibrium at price P_2 because households are unable to buy as much as they wish to at this price. Households would like to buy Q_2 tomatoes, but they cannot because at this price tomato producers are only willing to supply Q_1. The situation is now reversed compared to P_1. Tomato buyers are on the long side of the market and farmers and tomato sellers are on the short side. In this case, the difference between Q_2 and Q_1 is **excess demand** or unfulfilled demand. Households end up buying Q_1 tomatoes because this is the maximum quantity that tomato producers are prepared to sell at this price.

How a shift of supply disturbs market equilibrium

Once supply equals demand in a market — for example, at point X in Figure 3.17 — the market remains in equilibrium until an external event hits the market and causes either the market supply curve or the market demand curve to shift to a new position.

Figure 3.17 illustrates what happens in the tomato market when an event such as a bumper harvest causes the supply curve of tomatoes to shift rightwards, from S_1 to S_2. Before the shift of the supply curve,

P_1 was the equilibrium price of tomatoes. However, once the supply curve shifts, P_1 becomes a disequilibrium price. Too many tomatoes are offered for sale at this price, which means there is excess supply in the market. The excess supply is shown by $Q_2 - Q_1$, or the distance between X and V.

To get rid of this unsold stock, tomato producers reduce the price they are prepared to accept. The market price falls from P_1 to P_2, which eliminates the excess supply. In the new equilibrium, planned supply once again equals planned demand, but at the lower equilibrium price of P_2.

TEST YOURSELF 3.13
The equilibrium price for centre court tickets at an international tennis final is £5,000. The official ticket provider sells these tickets for £100. What do you think happens in the second-hand market for tickets to watch the tennis match?

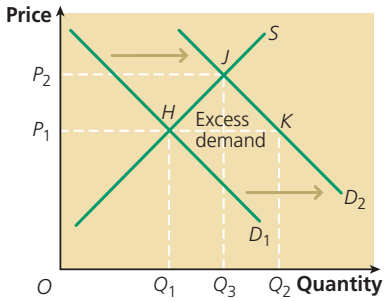


Figure 3.18 The effect of a rightward shift of the market demand curve for tomatoes

STUDY TIP
Make sure you can distinguish between a shift of a supply or demand curve, and the adjustment to a new equilibrium along the curve that does not shift.

How a shift of demand disturbs market equilibrium

Figure 3.18 shows what happens in the market for tomatoes following an increase in consumers' incomes. Tomatoes are usually considered a normal good: that is, a good for which demand increases as income increases. Before the increase in consumers' incomes, the equilibrium price of tomatoes was P_1, determined at the intersection of curves D_1 and S. At this price, planned demand equals planned supply. However, increased incomes shift the market demand curve rightwards from D_1 to D_2. Immediately, disequilibrium replaces equilibrium in the market. The rightward shift of demand creates excess demand in the market, as long as the price remains at P_1. Excess demand is shown by $Q_2 - Q_1$, or the distance between H and K.

The market adjustment mechanism now swings into action to get rid of the excess demand. The price increases to P_2 to eliminate the excess demand, and the quantity of tomatoes bought and sold rises to Q_3. In response to the increase in demand from H to K, there is a movement along the supply curve between H and J (an extension of supply) to establish the new equilibrium.

TEST YOURSELF 3.14
There are 30 million customers and 1 million firms producing the good in a particular market in the UK. Explain why you would classify this market as being competitive or uncompetitive.

SYNOPTIC LINK
See Figure 3.8 on page 59, which shows a leftward shift of a supply curve. This could be caused by the imposition of a specific tax or by a cut in a subsidy previously given to firms.

EXTENSION MATERIAL

How the effect of an expenditure tax depends on elasticity of demand

Figure 3.8, earlier in the chapter, can be used to show what can happen when an expenditure tax is imposed on a particular good. The tax shifts the good's supply curve to the left. From the point of view of the firms that produce and sell the good, the tax has the same effect as a rise in costs of production, such as a rise in wage costs. As is the case with cost increases, by raising the price of the good to cover the tax, firms try to increase the price charged to customers by the full amount of the tax. However, their ability to do this depends on price elasticity of demand for the good or service in question.

Figure 3.19 shows that when demand is relatively elastic, consumer resistance means that some, but not all, of a tax (in this case, a specific tax) is passed on to consumers as a price rise. The tax per unit (labelled T in Figure 3.19) is measured by the vertical distance between S_1 (the supply curve before the tax was imposed) and S_2 (the supply curve after the tax was imposed). Immediately after the imposition of the tax, firms may try to raise the price to $P_1 + T$, passing all the tax on to consumers. However, there is excess supply at this price. To get rid of the excess supply, the price falls to P_2. In the new equilibrium, part, but not all, of the tax has been passed onto consumers as a price rise.

The part of the tax passed on to consumers is called the *shifted incidence* of the tax. The rest of the tax (the *unshifted incidence*) is borne by firms or producers. In Figure 3.19, the total tax revenue paid by firms to the government is shown by the rectangle bounded by heavy black lines. The part of the tax rectangle above what was previously the equilibrium price (P_1), shows the shifted incidence of the tax. The part of the tax rectangle below P_1 shows the unshifted incidence.

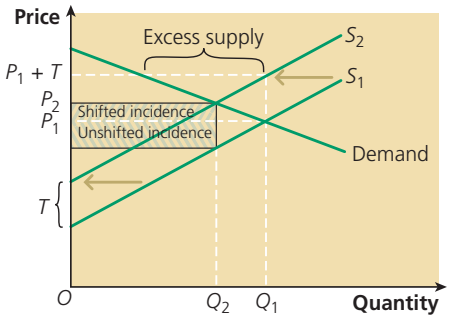


Figure 3.19 Shifting the incidence of a tax when demand is price elastic

You should now draw diagrams similar to Figure 3.19, but with perfectly elastic, relatively inelastic and completely inelastic demand curves. The diagrams will show that firms' ability to pass the incidence of a tax on to consumers as a price rise is greatest when demand is completely inelastic, and non-existent when demand is perfectly elastic.

Students often confuse the effect of an increase in an indirect tax imposed on firms with the effect of a direct tax such as income tax imposed on individuals. Whereas a tax imposed on firms shifts the *supply curve* of a good, income tax shifts the *demand curve* for a good by reducing consumers' incomes. An increase in income tax shifts the demand curve for normal goods leftwards, but if the good is an inferior good, the demand curve shifts rightwards.

Finally, note that subsidies granted to firms have the opposite effect to taxes imposed on them. Subsidies granted to firms shift the supply curve rightwards, showing that firms are prepared to supply more of the good at all prices.

TEST YOURSELF 3.15

Distinguish between an expenditure tax and an income tax. Illustrate on a supply and demand graph the possible effect of an increase in both the percentage rate at which VAT is levied on the good and the rate at which income tax is levied, assuming the good is an inferior good.

STUDY TIP

Many students never really get to grips with microeconomic analysis because they fail to understand the difference between market *plans* and market *action*. Your market plans are what you wish to do when you go shopping. Your market action is what you *end up doing*, i.e. the goods you actually purchase.

QUANTITATIVE SKILLS 3.3

Worked example: calculating the equilibrium price of a good

Table 3.2 shows the demand and supply schedules for chocolate bars.

Table 3.2 Demand and supply schedules for chocolate bars

Price per bar (£)	Quantity of bars demanded per week	Quantity of bars supplied per week
0.75	180	240
0.70	200	200
0.65	220	160
0.60	240	120

As a result of a fall in the price of cocoa beans, the supply of chocolate bars rises by 60 bars at all prices. What is the new equilibrium price of chocolate bars?

According to the table, the initial equilibrium price of chocolate bars is 70p, at which demand and supply are equal at 200 chocolate bars. If 60 more chocolate bars are supplied at each price, following the fall in the cost of manufacturing the bars, 300 bars are supplied at a price of 75p, 260 bars at a price of 70p, and 220 bars at a price of 65p. This is the new equilibrium price because at this price, demand equals supply at 220 bars. The supply curve has shifted upwards by 60 at each price.

APPLICATION OF ECONOMICS IN THE REAL WORLD 3.3

Auctions

In theory, an auction provides a quick and efficient method of establishing equilibrium in a market. Auctions have been brought into many people's everyday lives through sites such as eBay. But they also have a long history spanning many different domains. For example, the US government uses auctions to sell Treasury bills and timber and oil leases, 'auction houses' Christie's and Sotheby's use them to sell art, and Morrell & Co. and the Chicago Wine Company use them to sell wine.

Each bidder has an intrinsic value for the item being auctioned — the bidder is willing to purchase the item for a price up to this value, but not for any higher price.

Three types of auction at which a single item is sold are:

- **Ascending-bid auctions**, also called English auctions. The seller gradually raises the price, bidders drop out until only one bidder remains, and that bidder wins the object at this final price.
- **Descending-bid auctions**, also called Dutch auctions. The seller gradually lowers the price from a high initial value until the first moment when a bidder accepts and pays the current price. These auctions are called Dutch auctions because flowers have long been sold in the Netherlands using this procedure.
- **First-price sealed-bid auctions.** In this kind of auction, bidders submit simultaneous 'sealed bids' to the seller. The terminology comes from the original format for such auctions, in which bids were written down and provided in sealed envelopes to the seller, who would open them all together. The highest bidder wins the object and pays the value of their bid.

Follow-up questions

1 What is meant by 'equilibrium in a market'?
2 What type of auction are eBay auctions?
3 Research another example of a descending-bid auction.
4 In the UK, second-hand or used cars are sometimes sold at auction. Describe another way in which second-hand cars are sold.

SECTION 3.5 SUMMARY

- Market equilibrium occurs at the price at which the demand curve crosses the supply curve: that is, where demand equals supply.
- Disequilibrium occurs when there is either excess demand or excess supply in the market.
- In a competitive market, changes in the market price eliminate excess demand or excess supply; this is how the price mechanism helps to allocate scare resources.

3.6 The interrelationship between markets

So far in this chapter we have looked at how the price mechanism operates in a competitive market. We have seen how shifts of either the demand or supply curve for the good disturb market equilibrium and trigger an adjustment process to establish a new equilibrium.

Shifts of curves are often caused by events taking place in other markets in the economy. On the supply side, they can be caused by a change of price of a good in joint supply (see below). On the demand side, shifts can be caused by a change in price of a good in complementary demand (joint demand) or a substitute good, both of which we have already mentioned. They can also be caused by a change in the price of a good in composite demand or derived demand, as will be explained below.

Joint supply, joint demand, competitive demand, composite demand and derived demand

Joint supply

KEY TERM
joint supply when one good is produced, another good is also produced from the same raw materials, perhaps as a by-product.

Joint supply occurs when production of one good leads to the supply of a by-product. Suppose, for example, that the demand for beef increases, possibly because of rising incomes in economically developing countries. In these countries, meat is a relatively expensive luxury which people tend to buy more of if their incomes rise. The slaughter of more cows to meet this demand leads to production of more cow hides, which increases the supply of leather.

The interrelationship between the beef and leather markets is shown in Figure 3.20. Note that the price of beef *rises* following the rightward shift of the *demand* curve for beef, but the price of leather *falls* following the rightward shift of the *supply* curve of leather. A rise in the price of the first good leads to a shift of the supply curve of the other good in joint supply. In this example, beef is the main product and leather is the by-product, though the relationship could be reversed.

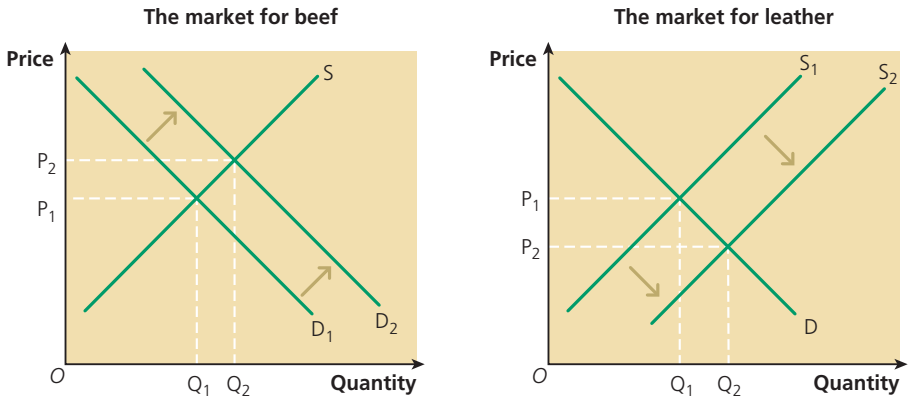


Figure 3.20 The interrelationship between two goods in joint supply

STUDY TIP

Make sure you don't confuse joint supply with *competing* supply, which is related to the concept of composite demand. The relationship between food and biofuel, shown in the Application of Economics in the Real World 3.4 which follows shortly, provides a topical example. Increased demand for biofuels such as ethanol has diverted crop growing away from food supply to the supply of fuel for motor vehicles. Because farmers producing crops such as wheat, maize and sugar can earn a higher price by selling their produce to energy companies, the supply curve of crops for food is shifting leftwards. We will leave it to you to draw appropriate diagrams to illustrate competing supply.

KEY TERMS

competitive demand when a good is viewed by consumers as an alternative for another good, i.e. the two goods are substitutes.

composite demand demand for a good which has more than one use, which means that an increase in demand for one use of the good reduces the supply of the good for an alternative use. It is related to the concept of competing supply.

derived demand demand for a good or factor of production, wanted not for its own sake, but as a consequence of the demand for something else.

TEST YOURSELF 3.16

With the help of two examples, distinguish between joint demand and composite demand.

Joint demand and competitive demand

As noted earlier, an increase in the price of a good in joint demand (or a complementary good) has the opposite effect to an increase in the price of a substitute good (or a good in **competitive demand**). For example, Sony games consoles and Sony games are in joint demand, but Sony and Xbox consoles are in competing demand, so they are substitute goods. Following a significant rise in the price of Sony consoles, demand for them falls, which in turn reduces the demand for Sony games. The demand curve for Sony games shifts leftwards. But the demand curve for Xbox consoles shifts rightwards, assuming that consumers consider an Xbox console to be a good substitute for a Sony console.

Composite demand

Students often confuse competitive demand, which occurs in the case of substitutes, with *composite demand* and *derived demand*. **Composite demand** is demand for a good which has more than one use. An increase in demand for one use of the good reduces the supply of the good for an alternative use: for example, if more wheat is used for biofuel, less is available for food, unless wheat growing increases.

Derived demand

By contrast, **derived demand** for a good occurs when a good is necessary for the production of other goods. The demand for capital goods such as machinery and raw materials is derived from the demand for consumer goods or finished goods. If the demand for cars falls, so does the demand for engines and gear boxes.

STUDY TIP

When you are studying interrelated markets, it is often useful to make use of the concept of cross elasticity of demand. This means you can comment on the extent to which two products are substitutes for, or complements to, each other.

APPLICATION OF ECONOMICS IN THE REAL WORLD 3.4

Composite demand and competing supply: biofuels and food

In recent years, high food prices have led experts to warn of the danger of a global food crisis. Many factors have contributed to the price rises, but the growth in production of biofuels has been one of the most important. Much of US maize production goes into biofuels. In 2022, 8.4% of motor fuels refined in the UK were biofuels, often made from wheat and maize, which are staple foods in the developing world.

Increased demand for biofuels inevitably drives food prices higher, though second- and third-generation biofuels are sourced in other ways: for example, from wood cellulose. Biofuel use is set to grow, and it is still the case that less food is grown as biofuel production increases.

Source: ONS

Follow-up questions

1 How are composite demand and competing supply related?
2 Explain how diverting crop production to meet the demand for biofuel is affecting world poverty.
3 Explain two causes, other than increased biofuel production, of recent changes in food prices.
4 In your view, are biofuels a source of clean energy or a source of dirty energy?

APPLICATION OF ECONOMICS IN THE REAL WORLD 3.5

Digital downloads and streaming replace CDs and DVDs

In 2000, when the first MP3 players were hitting the market, no one anticipated that sales of music downloads would overtake CD and DVD sales. Even when the iTunes store opened in 2003, Apple was only vying for a small market share.

Worldwide music streaming revenues are estimated to have risen, while sales of physical formats have dropped, yet within the physical format, sales of traditional vinyl records are staging a recovery.

What have we given up by adopting music streaming technology? Booklets, posters and CDs we can hold in our hands. What are we gaining? Instant satisfaction, convenience and mobile purchasing power.

As sales have moved online, many physical shops have closed. Now it seems that CDs could be phased out. How will our ways of consuming music evolve as time goes on?

Follow-up questions

1 Over the last 50 years, demand for recorded music switched first from vinyl records to CDs and then to downloading files and on to streamed music. However, vinyl sales are now growing again. Explain two reasons for these changes in demand.
2 How would you describe the demand relationship between CDs and music streaming?
3 Why have many music shops closed in recent years?
4 How have online firms such as Spotify, Google and Amazon affected UK retail markets in recent years?

EXTENSION MATERIAL

Why prices are often unstable in agricultural markets

In recent history, agricultural markets for foodstuffs and primary products such as rubber have experienced two closely related problems:

- Until recently, there was a long-run trend for agricultural prices to fall relative to those of manufactured goods.
- Prices have fluctuated considerably from year to year.

Agricultural markets are prone to disequilibrium and random shifts of the supply curve from year to year, caused by climatic factors, natural disasters and other world events, such as the Russian invasion of Ukraine in 2022. Climatic factors include long- and short-term changes in weather which can have significant impacts upon supply. These factors can all lead to unacceptable fluctuations in agricultural prices which, as Chapter 8 explains, may require government intervention to stabilise the price.

The long-run downward trend can be explained by shifts of the demand and supply curves for agricultural products over extended periods of time. This is shown in Figure 3.21, where the equilibrium price for an agricultural product in an early historical period is P_1. Over time, both the demand and supply curves have shifted rightwards. The shift in the demand curve was caused, for example, by rising incomes and population growth, while improved methods of farming increased supply. But for many farm products this shift of supply has greatly exceeded the shift of demand, resulting in a fall to the lower equilibrium price P_2.

Since the global recession ended in 2009, we might be seeing the beginning of a long-run trend for food price rises. Can you think of reasons why this might be happening?

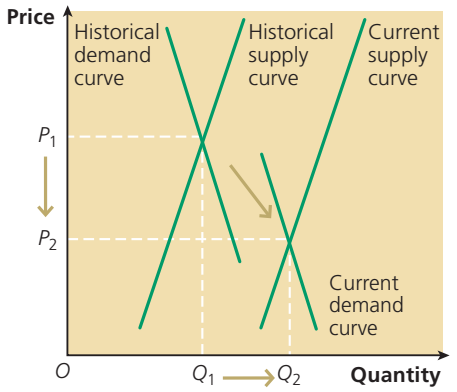


Figure 3.21 The long-run fall in the prices of agricultural products

Figure 3.22 provides an explanation of fluctuating farm prices. In the diagram, price volatility is caused by random shifts of the short-run supply curve in response to fluctuations in the harvest. The diagram shows two short-run supply curves: a 'good harvest' supply curve, S_1, and a 'bad harvest' supply curve, S_2. Weather conditions and other factors outside farmers' control shift the position of the supply curve from year to year between the limits set by S_1 and S_2. As a result, market prices fluctuate from year to year within the range of P_1 to P_2.

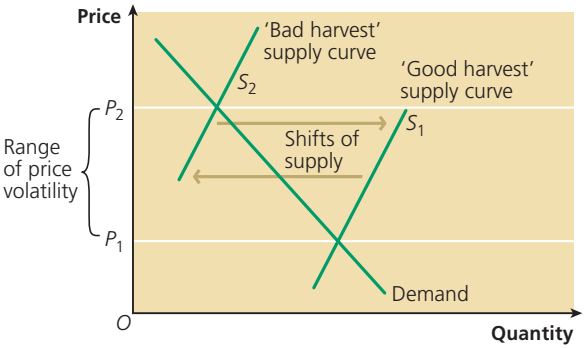


Figure 3.22 Fluctuating agricultural prices caused by shifts of supply

STUDY TIP

Make sure you understand the different ways in which governments can intervene in agricultural and commodity markets to try to stabilise prices.

QUANTITATIVE SKILLS 3.4

Worked example: calculating the income elasticity of demand for a good

Table 3.3 shows a price index for food covering the years 2018–22.

Table 3.3 Food price index, 2018–22

Year	Price index
2018	90
2019	100
2020	105
2021	120
2022	110

Which one of the following can be deduced from the data?

A Compared with 2021 the price of food was 10% cheaper in 2022.
B Farmers' revenue from the sale of food rose by 10% in 2019.
C The average price of food increased between 2018 and 2021.
D The price of food rose throughout the period from 2018 to 2022.

This question is testing your understanding of how index numbers are used in economics. Possible answers A, B and D are wrong, leaving C as the correct answer. A is wrong because food was 8.3% cheaper in 2022 not 10% cheaper. To calculate this, you take the change of index points of 10 between the two years as a ratio of 120 and turn this into a percentage. B is wrong because the index numbers do not allow you to calculate farmers' revenue. C is correct because the price index for food increased in every year between 2018 and 2021. D is wrong because the index did not increase between 2021 and 2022.

APPLICATION OF ECONOMICS IN THE REAL WORLD 3.6

Markets for new and second-hand cars

The number of new cars registered in 2021 in the UK rose by 1.0% compared with a year earlier, although sales were still 28.7% below 2019 levels. A car is a consumer durable good, delivering a constant stream of consumer services throughout its life. Provided it is properly looked after and escapes a serious crash, a new car typically lasts about 15 years. However, unlike a house, which is the ultimate consumer durable good (often with a much longer life than a person living in it), almost all new cars lose value or depreciate as soon as they have been driven off the showroom forecourt. This means that the prices of most cars fall throughout their lives.

In a recession or economic downturn, new-car owners may hang on to their cars for longer before they sell them — which decreases the supply of second-hand cars — and/or the demand for second-hand cars may fall because people cannot afford them. There are all sorts of possibilities.

Since 2017, the demand for new cars has been depressed for a number of reasons. First, there was bad publicity about 'dirty diesels'. Second, a new tougher emissions standard was introduced, which meant that owners of heavily-polluting cars pay more tax on their vehicles and on diesel fuel. Third, there was a lack of investment in the infrastructure needed for electrically powered vehicles. Fourth, consumer confidence fell following the Brexit referendum vote in 2016. Lastly, the Covid pandemic which was first detected in the UK in early 2020 further damaged consumer and business confidence as well as increasing the number of people working from home. The pandemic also affected global supply chains, driving up prices of car components.

Follow-up questions

1 What is a consumer durable good? Give another example besides a car.
2 Research what has happened to the prices of new cars in the UK between 2018 and the time you are reading this case study.
3 What is meant by depreciation and why do most cars depreciate?
4 How does the market for second-hand cars illustrate the problem of market misinformation?

QUANTITATIVE SKILLS 3.5

Worked example: calculating means and medians and constructing a line graph

Table 3.4 shows the sales of new vehicles in Great Britain between 2001 and 2021.

Table 3.4 New vehicle sales in Great Britain, 2001-21

Year	Sales (million)	Year	Sales (million)	Year	Sales (million)
2001	3.1	2010	2.4	2019	2.3
2002	3.2	2011	2.4	2020	1.6
2003	3.2	2012	2.5	2021	1.7
2004	3.2	2013	2.7		
2005	3.0	2014	3.0		
2006	2.9	2015	3.2		
2007	3.0	2016	3.3		
2008	2.7	2017	3.1		
2009	2.4	2018	2.4		

Source: Department of Transport

(a) Calculate the mean value to 2 decimal places, and the median values, of new vehicle sales in Great Britain between 2001 and 2021.

Mean values are calculated by adding up all the values in a data series and then dividing the total by the number of values in the series. The total of all the values is 57.3, which divided by 21 is 2.73 to 2 decimal places.

By contrast, the median value separates the higher half of the data series from the lower half. The data arranged from highest values to lowest values is 3.3; 3.2; 3.2; 3.2; 3.2; 3.1; 3.1; 3.0; 3.0; 3.0; 2.9; 2.7; 2.7; 2.5; 2.4; 2.4; 2.4; 2.4; 2.3; 1.7; 1.6. The highest values extend from 3.3 to 3.0 and the lowest values from 2.7 to 1.6. The median value is 2.9.

(b) Draw a continuous line graph from the data shown in the table.

The answer is shown in Figure 3.23.

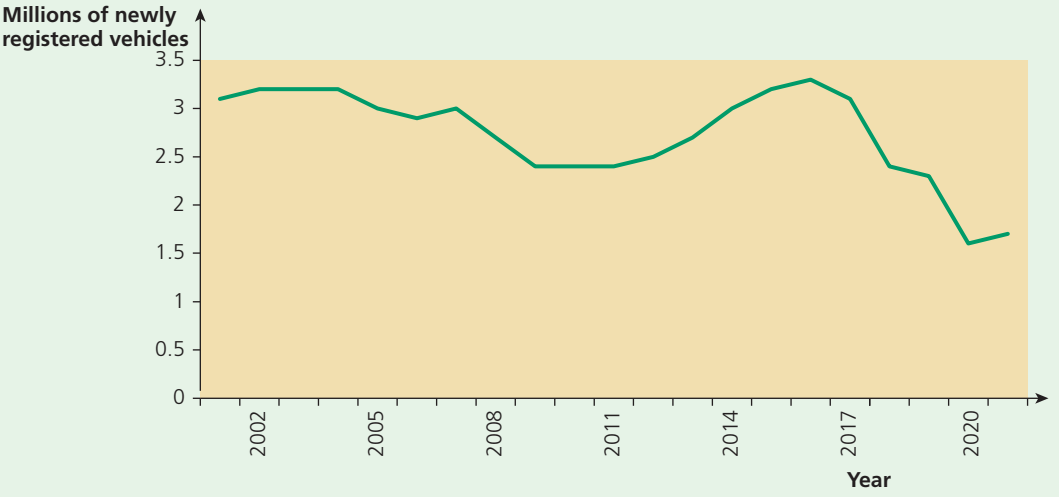


Figure 3.23 New vehicle sales in Great Britain, 2001-21

SECTION 3.6 SUMMARY

- You should practise applying market theory to different real-world markets.
- Make sure you understand the difference between joint demand and joint supply.
- Composite demand and derived demand are also relevant to the interrelationships between markets.
- It is important to understand the relationships between agricultural markets.
- Cross elasticity of demand is often an appropriate analytical tool when studying interrelated markets.

Questions

1 Evaluate the view that a fall in a good's price will inevitably lead to more demand for the good.

2 With an example of each, explain the difference between a normal good and an inferior good.

3 Explain how price elasticity of demand affects both total consumer spending and the total revenue received by firms when a good's price changes.

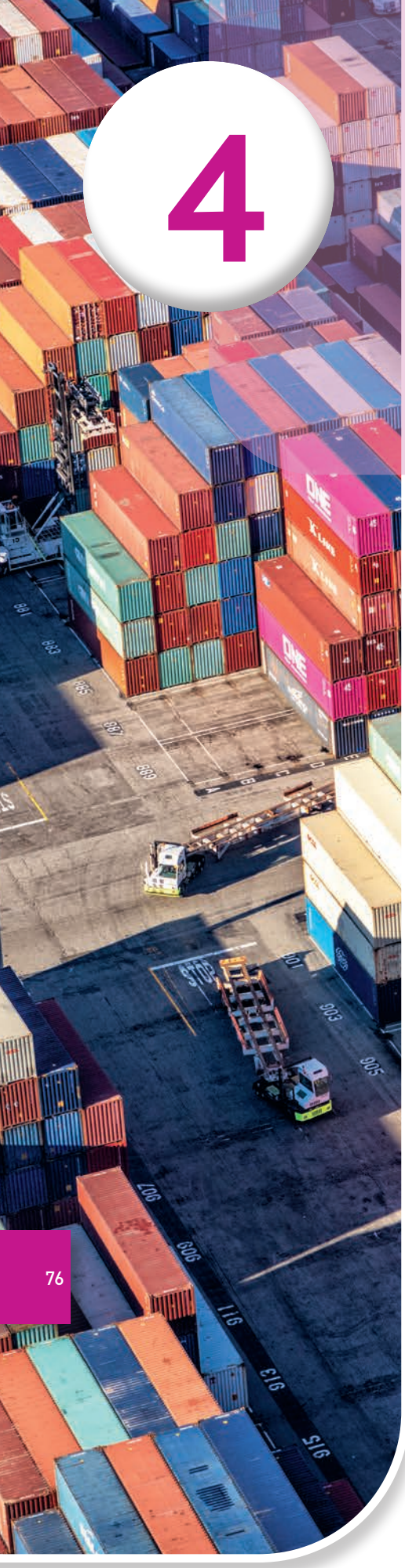
4 List and explain three factors which would cause a demand curve to shift to the left.

5 The cross elasticity of demand for good A with respect to the price of good B is +0. 8. Interpret this statistic.

6 List and explain three reasons why a supply curve may shift rightwards or downwards.

7 Explain how a low price elasticity of supply of new housing has affected UK house prices in recent decades.

8 With the help of an appropriate diagram, explain the effect on a good's price of a government subsidy granted to producers of the good.



4 Production, costs and revenue

This is the first of two chapters which cover an important part of microeconomic theory that economists often call the 'theory of the firm'. This chapter begins by looking at the nature of production, before going on to link production theory to the costs of production a firm incurs when it produces output. The chapter then explains the sales revenue a firm earns when it sells its output to customers, with the aim of making a profit.

The theory covered in this chapter leads seamlessly into Chapter 5, which explains the different market structures, such as perfect competition and monopoly, in which firms sell their output.

LEARNING OBJECTIVES

These are to understand:

- the meaning of production and the difference between production and productivity, especially labour productivity
- the concepts of specialisation and the division of labour
- how the laws of diminishing returns and returns to scale respectively help explain short-run and long-run production theory
- how short-run total, average and marginal cost curves are derived from short-run production theory
- how long-run cost curves are affected by the nature of returns to scale
- economies and diseconomies of scale in the economic long run
- the nature of the revenue curves facing a firm
- how revenue curves are dependent on the type of market structure in which the firm sells its output
- the role of profit in the economy
- how technological change can affect production and costs, and also competitiveness and market structure.

4.1 Production

Section 1.3 of Chapter 1 briefly explained the meaning of production and illustrated the basic nature of production in Figure 1.3. The chapter also mentioned the roles of factors of production such as labour, capital and enterprise or the entrepreneurial function in the production process. Entrepreneurs, such as Rihanna and Sir Richard Branson, are both decision-makers and the bearers of financial risk within a business.

TEST YOURSELF 4.1

Name four other important entrepreneurs besides Rihanna and Sir Richard Branson.

KEY TERMS

production converts inputs or factor services into outputs of goods and services.

factors of production inputs into the production process (e.g. land, labour, capital and enterprise).

Before proceeding any further with this chapter, refer back to Chapter 1 and read again what the chapter has to say on **production**.

Now look closely at Figure 4.1, which shows how production theory is the first of the five 'building blocks' of the theory of the firm.

The flow chart in Figure 4.1 shows that production theory leads into cost theory. Indeed, it is impossible to understand cost theory fully without first understanding the nature of production. The diagram also indicates that short-run cost theory derives from short-run production theory, and likewise long-run cost theory derives from long-run production theory.

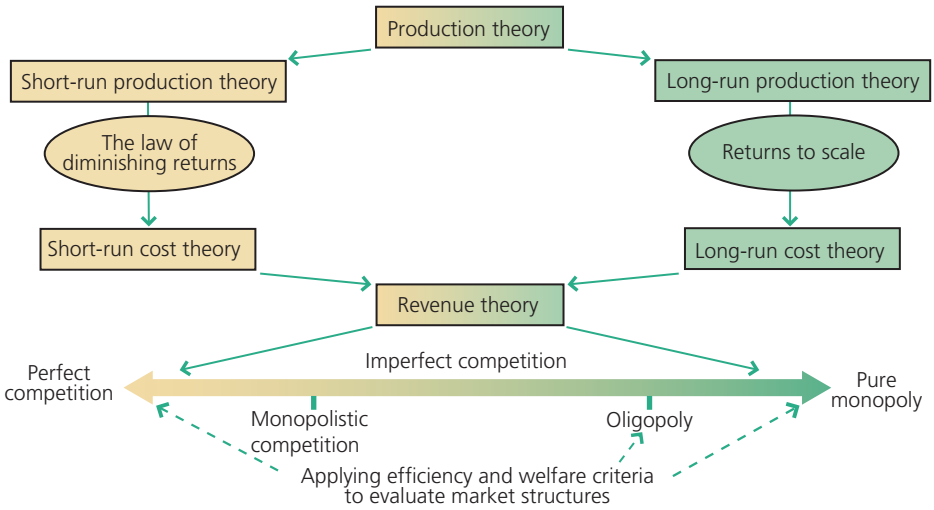


Figure 4.1 The 'building blocks' of the theory of the firm

Economists call the inputs into the production process the **factors of production**, which are listed in Figure 1.3 in Chapter 1. Four factors of production are usually identified. These are land, labour, capital and enterprise, the last often being called the entrepreneurial input. Land is literally the part of the Earth's crust which the firm owns or hires the services of. Labour includes all people employed by the firm who are paid wages or salaries. Capital comprises the capital goods which the firm owns or hires, together with intellectual capital it owns or hires.

Entrepreneurs are different from the other factors of production. They are the people who decide what to produce, how to produce and for whom to produce it. An entrepreneur decides how much of the other factors of production, including labour, to employ. The costs of employing land, labour and capital, together with the cost of the entrepreneur's own services, become the firm's costs of production. Essentially the entrepreneur is a financial risk-taker and decision-maker. Profit, which is the entrepreneur's financial reward, results from successful decision making. Entrepreneurial profit is the profit left over after the cost of employing the other factors of production is deducted from the sales revenue gained from the sale of the goods and services that the entrepreneur decides to produce.

KEY TERM

productivity gap the difference in labour productivity between, for example, the UK and other developed economies.

STUDY TIP

Productivity is a key concept in A-level economics. You must know the meaning of labour productivity, and also be aware of the meanings of other types of productivity. Be aware also of the UK's **productivity gap**, which is the difference in productivity levels between the UK and competitor countries. Productivity is also a key determinant of a country's international competitiveness.

The meaning of productivity, including labour productivity

KEY TERMS

productivity output per unit of input.

labour productivity output per worker.

capital productivity output per unit of capital.

Students often confuse productivity with production. While closely related, they do not have the same meaning. For most purposes, **productivity** usually means **labour productivity**, which is output per worker per period of time: for example, per week. However, **capital productivity** or output per unit of capital can also be measured, as can land productivity and entrepreneurial productivity. In reality, of course, all the employed factors of production contribute to both a firm's current level of output and any increase in the level of output.



Minis on the production line at the high-productivity BMW factory in Oxford

Labour productivity or output per worker is extremely significant in manufacturing industries, such as the car industry. In the 1990s and early 2000s, the Rover Car Group (which has since been bankrupted) struggled to survive in the UK car industry. Rover was unable to compete with Japanese car makers such as Nissan and Toyota. Nissan had invested in a then state-of-the-art factory near Sunderland. Labour productivity in the Rover factories amounted to only 33 cars per worker per year. By contrast, Nissan produced 98 cars per worker in its then brand-new factory. Given these figures, it is not surprising that the Rover Group, then owned by the German car manufacturer BMW, was forced to stop production of most of its cars in 2005. BMW continued to produce Minis in its high-productivity factories in Oxford.

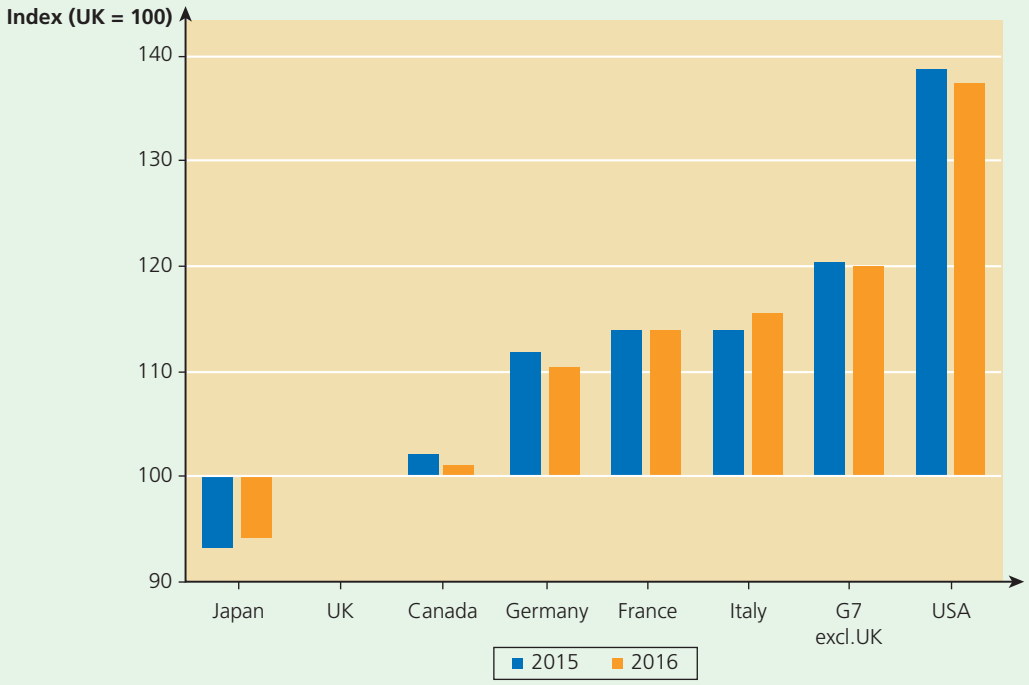
TEST YOURSELF 4.2

Explain the difference between production and productivity.

QUANTITATIVE SKILLS 4.1

Worked example: interpreting data presented in index number form

Figure 4.2 shows the gross domestic product (GDP) per hour in the G7 countries in 2015 and 2016.



Source: ONS

Figure 4.2 GDP per hour worked in G7 countries, 2015–16

(a) Who are the G7 countries?

As the horizontal axis of the graph indicates, the G7 or Group of Seven countries are Canada, France, Germany, Italy, Japan, the United Kingdom and the United States. G7 is an informal bloc of advanced or high-income industrialised democracies which meets annually to discuss issues such as global economic governance, international security and energy policy.

(b) Compare an index number with a percentage number.

Students often confuse data presented in index number form (as in Figure 4.2) with data presented in percentages. The confusion may arise because percentages add up to 100, and the number 100 is the number generally used for the base year in an economic index. (Note: the plural of 'index' is 'indices'.)

(c) Why are there no bars in the graph showing GDP per hour worked in the UK?

The data allow comparisons to be made between GDP per worker in each of the G7 countries. UK GDP per worker is depicted by the index number 100, which is shown by the horizontal line labelled 100. The index number for each of the other G7 countries is either higher or lower than 100. In Italy, for example, the 2015 index number of approximately 114 means that Italian labour productivity was about 14% higher than UK labour productivity, rising to approximately 16% higher in 2016.

The data are presented in index number form, which means we can compare *relative* levels and changes across the G7 countries, but not *absolute* levels and changes. We also cannot conclude from the data that Italian workers enjoy higher incomes and living standards than British workers. Although the data *infer* that living standards are higher in Italy than in the UK, this cannot be concluded definitely from the data given here. More evidence is needed to reach this conclusion.

The importance of productivity

In recent years a big problem has adversely affected UK economic performance: the failure of labour productivity to recover from a relatively low level, compared to other countries such as Germany and the USA, in the years following the 2009 recession. This has been called the 'productivity puzzle'. Why has the UK economy performed less well than competitor countries in increasing labour productivity?

Among the explanations of the productivity puzzle that have been put forward are: inadequate investment in new capital goods, relatively low wages in the UK economy, and employers 'hoarding' rather than laying off workers in the recession, which, with depressed output, inevitably means that labour productivity falls. With regard to the latter argument, the fall in labour productivity has helped employment in the UK in the short run, but the long-run consequences of low productivity growth may be much less favourable.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.1

The UK's productivity puzzle

When the UK entered recession during the financial crisis in 2008, labour productivity fell significantly. However, by 2014 UK labour productivity had not recovered to its pre-recessionary 2008 level, with British workers producing about a fifth less for every hour worked than workers in other leading nations. In some years, there was weak recovery, but in other years productivity continued to fall. The failure of productivity to recover from its recessionary low has been called the UK's 'productivity puzzle'.

The UK's 'flexible' labour markets have been very effective in delivering part-time and temporary work, at low cost to employers. A flexible labour market means that workers are able to adapt quickly and efficiently to changes in conditions of demand and supply. In 2014, up to 5.5 million people may have been working on 'zero hours' contracts (in which employers are not required to guarantee a certain number of hours to a worker). At the same time, under-employment (those who would like to work more hours, but cannot do so) has been at record levels. When faced with collapsing markets in the 2008–09 recession, employers — rather than reducing the number of people in work — often tried to cut the wages and hours of those working.

This then led to labour 'hoarding' — keeping workers in employment even though they are in effect producing little or no output. Employers were reluctant to fire skilled workers, believing that they would be difficult to re-recruit at a later point in the economic upturn. A 2012 survey of private-sector employers by the Chartered Institute of Personnel Development reported that close to a third of businesses had more staff than they needed to fulfil current orders. Most said they were anxious to retain their skills base. From an employer's point of view, it makes sense to let some workers stand idle to avoid delays in training new workers when demand recovers.

Following a recession, it takes time for firms facing declining demand to shed labour. It also takes time for new, fast-growing sectors to mop up the labour. During this period of flux, too many workers are concentrated in declining industries, with the result that economic output is less than it could be.

Related to this, low labour productivity has been blamed on so-called 'zombie' firms. A 'zombie' firm is an under-performing company, with low labour productivity, which survives because it generates just enough sales revenue to enable interest to be paid on its debts. After previous recessions, banks and other creditors refused to rescue such poorly performing companies, which were allowed to go to the wall. However, this did not happen when the 2008 recession ended in 2009. Very low interest rates or easy credit meant that zombie firms survived. (In 2014, however, the Bank of England dismissed the view that zombie firms provided a significant explanation of low labour productivity.)

What are the consequences of the productivity puzzle? One possibility is that even greater sacrifices may be demanded from those in work, and from those looking for work. If productivity is not improving, reduced labour costs per hour could be achieved by cutting hourly pay. In a 'race to the

bottom', UK workers would be pushed to match the lowest competing wages elsewhere in the world. Although the national minimum wage is meant to prevent this happening, there is plenty of evidence that rogue employers flout the law. A better possibility would be to boost productivity. This would require increased investment in newer, more efficient equipment and improved infrastructure. Investment spending on research and development (R&D) might lead to gains over time. However, none of this may happen. In 2014, investment in the UK economy was much lower than it had been in 2008.

Follow-up questions

1 Explain the meaning of a 'flexible' labour market.
2 Find out what has happened to labour productivity in the UK in the three years before you read this case study. Has labour productivity recovered?
3 How are 'rogue employers' related to the so-called 'race to the bottom'?
4 Why is raising labour productivity an important policy objective for the UK government?

What is a firm?

KEY TERM
firm a productive organisation which sells its output of goods and/or services commercially.

A **firm** is a business enterprise that either produces or deals in and exchanges goods or services. Unlike non-business productive organisations, such as many charities, firms are commercial, earning revenue to cover the production costs they incur.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.2

Japanese manufacturing methods and labour productivity

Until about 30 years ago, most car factories were chaotic places. Modern car factories, by contrast, are much calmer. The difference between the noisy, confused old factories and the smooth-flowing world of the modern ones is the Toyota Production System (TPS), first developed in the 1950s by the Japanese car company. Central to the Toyota Production System, now adopted by all mass car producers, is 'lean manufacturing'.

The aim of lean manufacturing is to combine the best of both craftwork and mass production. It uses less of each input: less labour, less machinery, less space, less time in designing products. Mass production concentrates on reducing defects to a tolerable level. Lean production seeks to eliminate all defects; if something goes wrong, the whole assembly line stops while the fault is identified and put right. An old car factory would have produced a complete afternoon's worth of cars with the same defect. In a lean factory, the mistake is quickly rectified so that the production of mechanically perfect cars can continue.

Lean manufacturing rejects the old idea of making things in huge batches, which requires holding large buffer stocks of materials and components between each stage of the production process. Now each stage of manufacturing performed in the factory is done on demand. The process eliminates waste by making only as much as is wanted at any given time. The change has greatly affected labour productivity.

Follow-up questions

1 Explain how the changes in methods of production mentioned in the passage are likely to have affected labour productivity and costs of production within manufacturing industries.
2 Distinguish between craftwork and mass production.
3 Another feature of Japanese manufacturing methods is known as 'just in time' (JIT). Find out what this means.
4 Research one other feature of Japanese manufacturing methods that increases labour productivity.

EXTENSION MATERIAL

Ronald Coase and the nature of a firm

Way back in 1937, Professor Ronald Coase, who much later in 1991 received the Nobel prize in economics for his insights, set out to explain why firms exist. His answer was that firms exist because they reduce transaction costs, such as search and information costs, bargaining costs, costs of keeping trade secrets, and policing and enforcement costs.

Coase then asked, 'Why then don't firms become bigger and bigger? Why isn't all world production carried on by a single big firm?' He gave two main reasons. 'First, as a firm gets larger, there may be decreasing returns to the entrepreneurial function, that is, the costs of organising additional transactions within the firm may rise. Secondly, as the transactions which are organised increase, the entrepreneur fails to make the best use of the factors of production.' At a certain point, the gains from economies of scale are defeated by the costs of bureaucracy.

For further information on Coase and the nature of the firm, and on the different views of later economists, access on the internet the article by Steve Denning, 'Did Ronald Coase get economics wrong?', published in *Forbes* magazine on 25 September 2013.

SECTION 4.1 SUMMARY

- The 'building blocks' of the theory of the firm include production theory, cost theory and revenue theory.
- The key concept in short-run production theory is the law of diminishing returns, also known as the law of diminishing marginal productivity.
- Productivity, which is a key economic concept, is output per unit of input.
- Labour productivity is output per worker.
- Avoid confusing productivity with production. Production is a process or set of processes which converts inputs into outputs.
- A firm is a business enterprise that either produces or deals in and exchanges goods or services.

4.2 Specialisation, division of labour and exchange

The benefits of specialisation and division of labour

KEY TERMS

specialisation a worker only performing one task or a narrow range of tasks. Also, different firms specialising in producing different goods or services.

division of labour different workers performing different tasks in the course of producing a good or service.

Nearly 250 years ago, the economist Adam Smith first explained how, within a single production unit or firm (he took the example of a pin factory), output could be increased if workers specialised at different tasks in the manufacturing process.

Smith had established one of the most fundamental of all economic principles: the benefits of **specialisation** or the **division of labour**. According to Adam Smith, there are three main reasons why a factory's total output can be increased if workers perform specialist tasks rather than if each worker attempts all the tasks him- or herself. These are as follows:

- A worker will not need to switch between tasks, so time will be saved.
- More and better machinery or capital can be employed. (Employing 'more of the same' capital is called 'capital widening', while investing in 'state-of-the-art' new technology is called 'capital deepening'.)

- The 'practice makes perfect' argument that workers become more efficient or productive at the task they are doing, the longer they spend on the specialist task. However, this advantage can easily become a disadvantage if it involves 'de-skilling' and the development of boredom and alienation among workers.

Trade and exchange

KEY TERMS

trade the buying and selling of goods and/or services.

exchange to give something in return for something else received. Money is a medium of exchange.

For specialisation to be economically worthwhile for those taking part in the division of labour, a system of **trade** and **exchange** is necessary. This is because workers who completely specialise cannot enjoy a reasonable standard of living if forced to consume only what they produce. The obvious solution is to produce more than the worker actually needs, and then to trade the surplus for that produced by others.

Several hundred years ago, people living in rural communities in the UK could specialise and then trade whatever they produced through barter. Thus, a farmer might harvest wheat, part of which was then exchanged for services provided by local grain millers and village blacksmiths.

But successful barter requires a 'double coincidence of wants'. Not only must the farmer require the services of the blacksmith; the blacksmith must want the wheat produced by the farmer, and a rate of exchange must be agreed for the two products. As this example suggests, it is reasonably easy to achieve the double coincidence of wants in a small community where people live close to each other and where only a few goods and services are produced and exchanged. However, in modern economies in which a vast number of goods are produced, reliance on barter holds back the growth of the economy. In such economies, barter is an extremely inefficient method of exchange.

These days, when we buy or sell a good or a service, we almost always use money. We finance the transaction either with cash or with a debit card or cheque drawn on a bank or building society deposit. Using money is much more efficient than bartering, as there is no need for a double coincidence of wants. Suppose I want to buy a television set and that I also have a second-hand car I wish to sell. Assuming you wish to buy a second-hand car and have an old TV set you want to get rid of, if we barter the goods, you must want my car and I must want your TV set. We must also agree that the two goods have the same value. But if we use money rather than barter, you pay for my car with money, which I can then use to buy a TV set or whatever I want from somebody else. I could also save the money rather than spend it. Used in this way, money enables the economy to achieve much greater specialisation and division of labour than is possible with barter.

However, we are increasingly moving towards a 'cashless society' in which credit cards, debit cards and contactless smartphone payments are replacing coins and notes as means of exchange. Some economists even believe that so-called cryptocurrencies such as the Bitcoin will replace all current forms of money, though at present this is a minority view. Refer to Chapter 12 on financial markets and monetary policy to find out more about the various forms of money and means of payment.



Our society is becoming increasingly cashless

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.3

Adam Smith's pin factory

Adam Smith was an eighteenth-century Scottish philosophy professor, and later customs commissioner, who is often said to be the founder of modern economics. In his book *An Inquiry into the Nature and Causes of the Wealth of Nations*, which was published in 1776, Adam Smith used the example of a local pin factory to explain how the division of labour among workers greatly increases their ability to produce. Here is a slightly abridged version of what Adam Smith wrote:

> A workman not educated in the business of pin making could scarce, perhaps, with his utmost industry, make one pin in a day, and certainly could not make twenty. But in the way in which this business is now carried on, one man draws out the wire, another straights it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the pin head. The business of making a pin is divided into about eighteen distinct operations. Ten persons could make among them upwards of forty-eight thousand pins in a day. Each person, therefore, making a tenth part of forty-eight thousand pins, might be considered as making four thousand eight hundred pins in a day. But if they had all wrought separately and independently, and without any of them having been educated to this peculiar business, they certainly could not each of them have made twenty, perhaps not one pin in a day.
>
> This great increase in the quantity of work is a consequence of the division of labour. There are three different aspects of this: first, the increase of dexterity in every particular workman; secondly, the saving of the time which is commonly lost in passing from one species of work to another; and lastly, the invention of a great number of machines which enable one man to do the work of many.

Follow-up questions

1. What is meant by the 'division of labour'?
2. What effects does the division of labour have on production and costs?
3. Describe another example, besides pin factories, of firms benefiting from the division of labour.
4. According to Adam Smith, 'This great increase in the quantity of work is a consequence of the division of labour.' Explain one way in which this has affected the size of firms.

SECTION 4.2 SUMMARY

- Specialisation occurs in more than one context. These include groups of workers specialising in producing different goods or services, and different workers specialising in different tasks, such as cleaners and managers within a place of work.
- With the division of labour, different workers specialise in different tasks in the course of producing a good or service.
- Trade is the buying and selling of goods and/or services.
- Exchange means to give something in return for something else received.

4.3 The law of diminishing returns and returns to scale

The short run and the long run

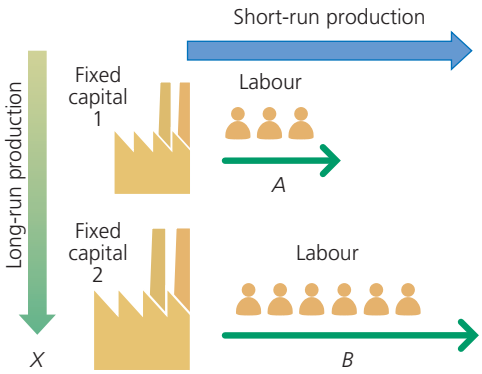


Figure 4.3 Short-run and long-run production

In microeconomic theory, the short run is defined as the time period in which at least one of the factors of production is fixed and cannot be varied. This means that the only way in which a firm can produce more in the short run is by adding more variable factors to the fixed factors of production.

In Figure 4.3, the horizontal arrows labelled *A* and *B* show short-run production taking place. By contrast, the movement along the vertical arrow *X* depicts the long run, defined as the time period in which the *scale* of all the factors of production can be changed.

The law of diminishing returns

The law of diminishing returns is a short-run law which states that as a variable factor of production is added to a fixed factor of production, eventually both the marginal returns and then the average returns to the variable factor of production begin to fall.

KEY TERM
marginal returns of labour the change in the quantity of total output resulting from the employment of one more worker, holding all the other factors of production fixed.

Table 4.1 shows what might happen in a small musical instrument workshop assembling guitars when the number of workers employed increases from zero to 10. The first worker employed assembles 1 guitar a day, and the second and third workers respectively add 7 and 10 guitars to the workshop's total daily output. These figures measure the marginal returns (or marginal product) of each of the first three workers employed. The **marginal returns of labour** are the addition to total output brought about by adding one more worker to the labour force.

In Table 4.1, the first five workers benefit from increasing marginal returns (or increasing marginal productivity). Each additional worker increases total output by more than the amount added by the previous worker. Increasing marginal returns are very likely when the labour force is small. In this situation, employing an extra worker allows the workforce to be organised more efficiently. By dividing the various tasks of production among a greater number of workers, the firm benefits from specialisation and the division of labour. Workers become better and more efficient in performing the particular

Table 4.1 Short-run guitar production with fixed capital

	Variable labour										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Total returns	0	1	8	18	32	50	64	70	72	68	60
Average returns	–	1	4	6	8	10	10. 7**	10	9	7.6	6
Marginal returns	1	7	10	14	18*	14	6	2	–2	–8	

* The point of diminishing marginal returns

** The point of diminishing average returns

Note: Total, average and marginal returns are often called total, average and marginal product. For example, in Table 4.1 the 'marginal returns' of labour can be called the 'marginal product' of labour.

tasks in which they specialise, and time is saved that otherwise would be lost as a result of workers switching between tasks.

But as the firm adds labour to fixed capital, eventually the **law of diminishing marginal returns** (or **law of diminishing marginal productivity**) sets in. In this example, the law sets in when the sixth worker is employed. The marginal return of the fifth worker is 18 guitars, but the sixth worker adds only 14 guitars to total output. Diminishing marginal returns set in because labour is being added to fixed capital. When more and more labour is added to fixed plant and machinery, eventually workers begin to get in each other's way and the marginal returns of labour fall, though not often at a labour force as small as six workers.

Note that the impact of diminishing marginal returns does not mean that an extra worker joining the labour force is any less hardworking or motivated than their predecessors. (In microeconomic theory, we often assume that workers and other factors of production are completely interchangeable and homogeneous.) In the economic short run, as more labour is added to a fixed amount of capital or machinery, the possibilities for further specialisation and division of labour eventually become exhausted.

KEY TERM

law of diminishing marginal returns (also known as **law of diminishing marginal productivity**) a *short-term* law which states that, as a variable factor of production is added to a fixed factor of production, both the marginal and eventually the average returns to the variable factor will begin to fall.

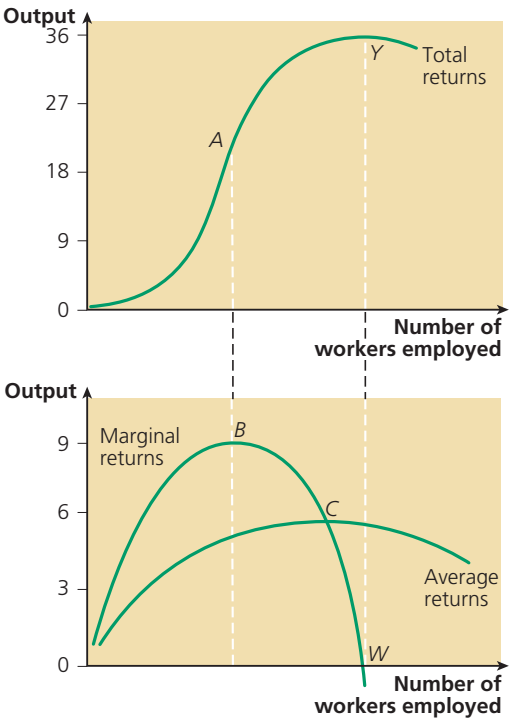


Figure 4.4 Total, marginal and average returns curves

The law of diminishing returns shown on a diagram

The two panels of Figure 4.4 illustrate the law of diminishing returns. In the upper panel of the diagram, diminishing marginal returns begin to operate at point *A*. Up to this point, the slope of the **total returns** or total product curve increases, moving from point to point up the curve. This shows the labour force benefiting from increasing marginal returns. When diminishing marginal returns set in, the total returns curve continues to rise as more workers are combined with capital, but the curve becomes less steep from point to point up the curve. Point *Y* shows where total returns begin to fall. Beyond this point, additional workers begin to get in the way of other workers, so the marginal returns to labour become negative.

KEY TERM

total returns the whole output produced by all the factors of production, including labour, employed by a firm.

KEY TERMS

average returns of labour total output divided by the total number of workers employed.

In the lower panel of Figure 4.4, the law of diminishing returns sets in at point *B*, at the highest point on the marginal returns curve. To the left of this point, increasing marginal returns are shown by the rising (or positively sloped) marginal returns curve, while beyond this point, diminishing marginal returns are depicted by the falling (or negatively sloped) marginal returns curve. Likewise, the point of diminishing **average returns of labour** is located at the highest point of the average returns curve at point *C*. Finally, marginal returns become negative beyond point *W*.

It is important to understand that all three curves in Figure 4.4 (and the three rows of Table 4.1) contain the same information, but the information is used differently in each curve (and row). The total returns curve plots the information *cumulatively*, adding the marginal returns of the last worker employed to the total returns before the worker joined the labour force. By contrast, the marginal returns curve plots the same information *non-cumulatively*, or as separate observations. Finally, at each level of employment, the average returns curve shows the total returns of the labour force divided by the number of workers employed.

STUDY TIP

Negative marginal returns are not a result of workers being stubborn or lazy. Neither are they because the first workers employed are more efficient than those who are employed later. As we said earlier, in the abstract world of microeconomic theory, workers are treated as equally able, homogeneous units.

EXTENSION MATERIAL

The relationship between marginal returns and average returns

The relationship between the marginal returns of labour and the average returns of labour is an example of a more general relationship that you need to know. (Shortly, we will provide a second example, namely the relationship between marginal costs and average costs of production.)

Marginal and average curves plotted from the same set of data always display the following relationship:

- When the marginal is greater than the average, the average rises.
- When the marginal is less than the average, the average falls.
- When the marginal equals the average, the average is constant, neither rising nor falling.

It is vital to understand this relationship. It does *not* state that an average value will rise when a marginal value is rising; nor does it state that an average value will fall when a marginal value falls. As we saw in Figure 4.4, marginal returns begin to fall as soon as the law of diminishing marginal returns sets in. Nevertheless, as long as marginal returns are greater than the average returns of labour, the latter continue to rise. When marginal returns exceed average returns, the average returns curve is 'pulled up', even when the marginal returns curve is falling. But when the marginal returns curve cuts through the average returns curve (at point *C* in Figure 4.4), beyond that point the average returns of labour begin to fall. The marginal returns curve cuts through the average returns curve at the latter's highest point. Beyond this point, the marginal returns curve continues to fall, and because marginal returns are less than average returns, they 'pull down' the average returns curve.

The difference between marginal and average returns

Whereas marginal returns are the addition to total output attributable to taking on the last worker added to the labour force, the average returns at any level of employment are measured by dividing the total output of the labour force by the number of workers employed. The average returns of the labour force employed in the guitar workshop are shown by the middle row of data in Table 4.1. Note that in the table, the point of diminishing *average* returns occurs after the sixth worker is taken on, whereas diminishing marginal returns set in after the fifth worker is employed. The relationship between the marginal returns and the average returns of labour is illustrated in the lower panel of Figure 4.4.

QUANTITATIVE SKILLS 4.2

Worked example: an example of the relationship between marginal, average and total values

A firm has a fixed amount of capital and land, and increases output by employing additional labour according to the schedule in Table 4.2.

Table 4.2 Diminishing returns to labour

Labour	Output
1	20
2	42
3	68
4	93
5	100
6	90

(a) At what level of output do diminishing marginal returns to labour set in?

Diminishing marginal returns set in when the marginal return or marginal productivity falls for an extra worker added to the labour force. The marginal returns of the second worker, which are calculated by subtracting total output when only 1 worker is employed from total output when 2 workers are employed, are 22 units of output. Via a similar calculation, the marginal returns of the third worker are 26 units of output. However, diminishing marginal returns set in when a fourth worker is added to the labour force. The marginal returns of the fourth worker are 25 units of output.

(b) At what level of output do diminishing average returns to labour set in?

For each size of labour force, average returns are calculated by dividing total output by the number of workers employed. Average returns when the labour force is 1, 2, 3 and 4 workers are respectively outputs of 20, 21, 22.67 and 23.25, showing increasing average returns. However, when the fifth worker is added to the labour force, marginal returns fall to 7 units of output and average returns fall to 20 units of output. Diminishing *average* returns (falling average output per worker) have now set in.

(c) At what level of output do diminishing total returns to labour set in?

Diminishing *total* returns set in when the addition of an extra worker causes total output to fall. This happens when the sixth worker is added to the labour force. Note that marginal returns are now negative (–10 units of output). The workers are getting in each other's way to such an extent that total output falls.

STUDY TIP

In production theory, students often confuse the law of diminishing returns, which is a short-run law applying when at least one factor of production is fixed, with returns to scale, which relate to the long run when firms can change the scale of all the factors of production. You must avoid this mistake. As we explain shortly, the law of diminishing returns is important for explaining the shape of short-run cost curves, and likewise, returns to scale help to explain the shape of long-run cost curves and the concepts of economies and diseconomies of scale.

Returns to scale

KEY TERMS

returns to scale the rate by which output changes if the scale of all the factors of production is changed.

plant an establishment, such as a factory, a workshop or a retail outlet, owned and operated by a firm.

Figure 4.3 earlier in the chapter illustrates the important distinction between returns to a variable factor of production, which occur in the short run, and **returns to scale**, which operate only in the economic long run. Suppose that a firm's fixed capital is represented by **plant** size 1 in the diagram.

Initially, the firm can increase production in the short run, by moving along the horizontal arrow *A*, employing more variable factors of production such as labour. To escape the impact of short-run diminishing marginal returns which eventually set in, the firm may make the long-run decision to invest in a larger production plant, such as plant size 2. The movement from plant size 1 to plant size 2 is shown by the movement along the vertical arrow *X* in the diagram. Once plant size 2 is in operation, the firm is in a new short-run situation, able to increase output by moving along arrow *B*. However, the impact of short-run diminishing returns may eventually cause the firm to expand the scale of its operations again in the long run.

Contrasting increasing, constant and decreasing returns to scale

KEY TERMS

increasing returns to scale when the scale of all the factors of production employed increases, output increases at a faster rate.

constant returns to scale when the scale of all the factors of production employed increases, output increases at the same rate.

decreasing returns to scale when the scale of all the factors of production employed increases, output increases at a slower rate.

It is important to avoid confusing returns to scale, which occur in the long run when the scale of *all* the factors of production can be altered, with the short-run returns that occur when at least one factor is fixed. With returns to scale there are three possibilities:

- **Increasing returns to scale**. If an increase in the scale of all the factors of production causes a more than proportionate increase in output, there are increasing returns to scale.
- **Constant returns to scale**. If an increase in the scale of all the factors of production causes the same proportionate increase in output, there are constant returns to scale.
- **Decreasing returns to scale**. If an increase in the scale of all the factors of production causes a less than proportionate increase in output, there are decreasing (or diminishing) returns to scale.

SECTION 4.3 SUMMARY

- The key concept in short-run production theory is the law of diminishing returns.
- The law states that as a variable factor of production is added to a fixed factor of production, eventually both the marginal returns and then the average returns to the variable factor of production begin to fall.
- The key concept in long-run production theory is returns to scale.
- Increasing returns to scale are likely to lead to economies of scale, which are defined as falling long-run average costs, as output increases.
- Decreasing returns to scale are likely to lead to diseconomies of scale, which are defined as rising long-run average costs, as output increases.

4.4 Costs of production

Economics students often confuse production and costs. Production, as previously explained, simply converts inputs into outputs, without considering the money cost of using inputs such as capital and labour. This section of the chapter explains, first, how short-run cost curves are derived from the short-run production theory we have already explained, and second, in a similar way, how long-run cost curves are derived from long-run production theory.

Short-run costs

The difference between short-run and long-run costs

We have already defined the short run as the time period in which at least one factor of production is fixed. Short-run costs are made up of fixed costs and the costs incurred when hiring the services of the variable factors of production. These are called variable costs of production. Written as equations:

total cost = total fixed cost + total variable cost

or:

$TC = TFC + TVC$

and:

average total cost = average fixed cost + average variable cost

or:

$ATC = AFC + AVC$

Long-run costs, by contrast, are always variable costs. Fixed costs don't exist in the economic long run when, as Figure 4.3 showed, a firm can move seamlessly from one size of plant to another.



Raw materials, such as coffee beans bought by a coffee producer, are a variable cost

KEY TERMS

fixed cost a cost of production which, in the short run, does not change with output.

variable cost a cost of production which changes with the amount that is produced, even in the short run.

total cost all the cost incurred when producing a particular size of output.

The difference between fixed and variable costs

In the *short run*, when the inputs divide into fixed and variable factors of production, the costs of production can likewise be divided into fixed and variable costs. **Fixed costs** are the costs a firm incurs when hiring or paying for the fixed factors of production. In the short run, capital is usually assumed to be a fixed factor of production, giving rise to costs which are unchanged in the short run. These include the cost of maintaining a firm's buildings, as well as the initial cost of acquiring buildings such as factory space and offices. By contrast, as the name implies, **variable costs** change as the level of output changes. The costs of hiring labour and buying raw materials are usually regarded as variable costs of production. It is worth remembering that in the *long run*, all costs are variable because the long run is defined as the time period in which all the factors of production can be changed.

The difference between total cost, average variable cost and marginal cost

When a firm increases its output, the **total cost** of production increases. At any level of output, the average cost (or unit cost) is calculated by

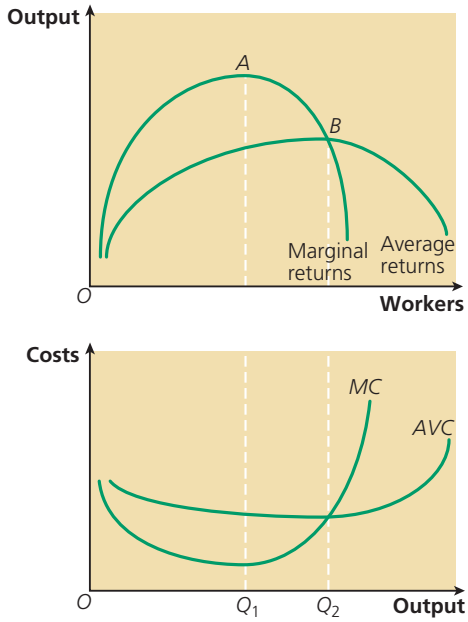


Figure 4.5 Deriving the marginal cost (*MC*) curve and the average variable cost (*AVC*) curve from the marginal and average returns curves

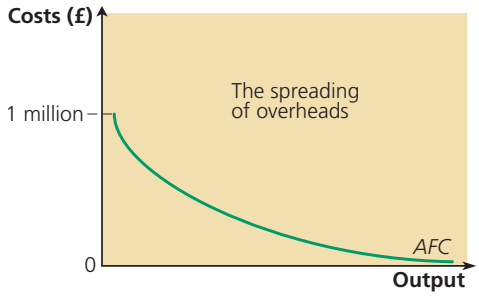


Figure 4.6 A firm's short-run average fixed cost (*AFC*) curve

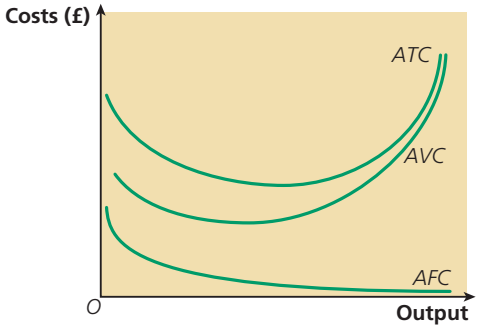


Figure 4.7 Deriving a firm's short-run average total cost (*ATC*) curve

dividing the firm's total cost of production by the size of output produced. An **average variable cost** curve is shown in the lower panel of Figure 4.5, which also shows the firm's **marginal cost** curve. Marginal cost is the addition to total cost resulting from producing one additional unit of output. We will shortly explain the shapes of all these cost curves.

KEY TERMS

average variable cost total variable cost divided by the size of output.

marginal cost addition to total cost resulting from producing one additional unit of output.

TEST YOURSELF 4.3

What is the difference between marginal costs and marginal returns?

The short-run average fixed cost (*AFC*) curve

Fixed costs of production are overheads, such as the rent on land and the maintenance costs of buildings, which a firm must pay in the short run. Suppose, for example, that a car manufacturing company incurs overheads of £1 million a year from an assembly plant it operates. If the plant only managed to produce one car a year, **average fixed costs** per car would be £1 million — the single car would bear all the overheads. But if the company were to increase production, average fixed costs would fall to £500,000 when two cars are produced, £333,333 when three cars are produced and so on. The firm's average fixed costs per unit of output fall as output increases, since overheads are spread over a larger output. *AFC* curves *always* slope downwards to the right, as shown in Figure 4.6, with average fixed costs approaching zero at very high levels of output, but never quite equalling zero.

The short-run average total cost (*ATC*) curve

The **average total cost** (*ATC*) curve shown in Figure 4.7 is obtained by adding together at each level of output the *AFC* and *AVC* curves. In the diagram, the firm's average costs of production initially fall as the size of output increases. However, for higher levels of output, beyond the lowest point on the *ATC* curve, average costs usually rise, leading to a U-shaped *ATC* curve.

KEY TERMS

average fixed cost total cost of employing the fixed factors of production to produce a particular level of output, divided by the size of output: $AFC = TFC \div Q$.

average total cost (also known as **average cost**) total cost of producing a particular level of output, divided by the size of output: $ATC = AFC + AVC$.

STUDY TIP

Make sure you practise calculating and plotting average and total costs from given data. This will help you to appreciate the shapes of, and interrelationships between, the various short-run cost curves.

QUANTITATIVE SKILLS 4.3

Objective test question: calculating average and total costs from given data

You should be able to calculate average and total costs from given data. This example takes the form of an objective test question in which the data are presented on a graph.

Figure 4.8 shows a firm's average variable costs and average fixed costs of production.

The total cost of producing 4 units is:

A £12 **B £48**
C £24 **D £72**

To arrive at the correct answer, which is D, two calculations are required. The first calculation involves adding *AVC* (£12) to *AFC* (£6) at an output of 4 units to arrive at average total cost (*ATC*), which is £18. The second calculation is multiplying *ATC* (£18) by output (4) to arrive at total cost, which is £72.

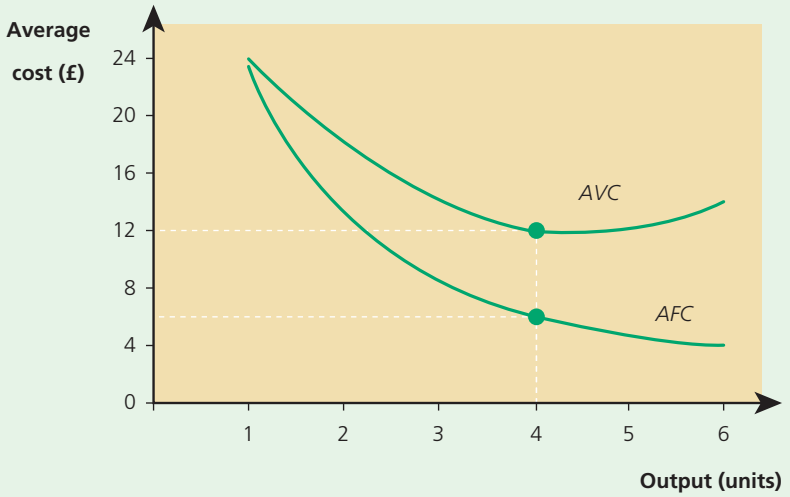


Figure 4.8 The average variable costs (*AVC*) and average fixed costs (*AFC*) of production for a firm

Bringing together the various short-run cost curves

Figure 4.9(a) is similar to Figure 4.7, but with the marginal cost curve added. As we explained earlier, by adding together the *AFC* and *AVC* curves, we get the *ATC* curve. The marginal cost curve cuts through from below *both* the *AVC* and *ATC* curves, at the lowest points of these curves. This is a further example of the mathematical relationship between the marginal and average values of a variable, which we explained on page 88.

In Figure 4.9(b), the *AFC* and *AVC* curves have been left out of the diagram in order to emphasise that the *MC* curve cuts through the lowest point of the U-shaped *ATC* curve.

STUDY TIP
Figure 4.9(b) is the diagram most often used by exam candidates when writing about short-run cost curves.

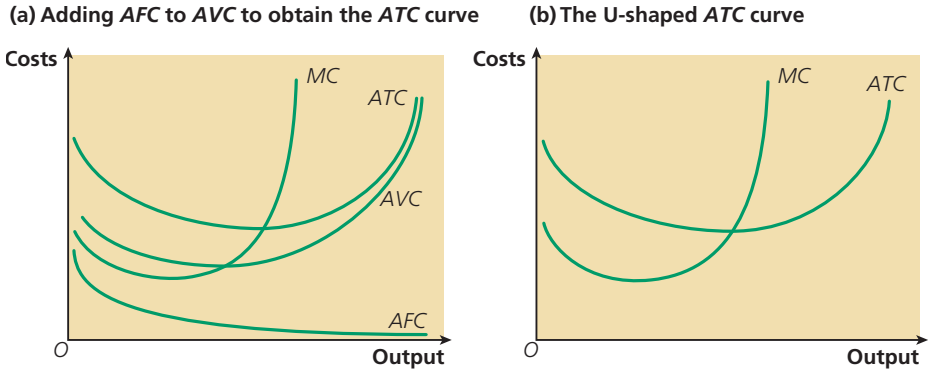


Figure 4.9 The relationships between marginal cost, average fixed cost, average variable cost and average total cost

How factor prices and productivity affect costs and choice of factor inputs

Factor prices are the prices that a firm pays for hiring the different factors of production (capital, labour etc.). Interest is the factor reward of capital, and wages and salaries are the factor reward of labour. We explained earlier in the chapter that capital productivity is output per unit of capital employed, and labour productivity is output per worker.

Economists generally assume that firms have a single business objective, namely to produce the level of output at which profit can be maximised. The profit-maximising level of output must, of course, be produced at the lowest possible cost for *that particular* level of output. This in turn means that the firm must employ the optimal factor combination when minimising the cost of producing the profit-maximising level of output.

This is where factor prices and productivity come in. If wage levels are low but the price of capital is high, then, provided we ignore labour and capital productivity, labour is more attractive than capital to employ. This suggests that a firm should use a labour-intensive method of production (lots of labour but little capital). Similarly, if labour is expensive to employ but capital is cheap, then cost minimisation is likely to require a capital-intensive method of production (lots of capital but little labour).

This becomes more complicated, however, when we introduce differences in labour and capital productivity into the decision-making process. For example, if wage levels are low but labour productivity is also low, a capital-intensive method of production can be preferable to labour-intensive production, if the high price of capital is more than offset by high capital productivity.

Capital-intensive production is often accompanied by high wage rates — though with relatively few workers employed. However, if consumers, both at home and abroad, demand more and more of the output produced by capital-intensive firms, employment of labour will also increase. Companies such as Nissan and Toyota, which employ robots rather than workers to produce parts of cars, have provided plenty of evidence of how the greater use of robotised machinery can enable both wage rates and employment to rise in the UK car industry.

QUANTITATIVE SKILLS 4.4

Objective test question: calculating average variable cost from given information on total cost

This is a similar objective test question to the previous one, but this calculation works in the opposite direction, asking you to calculate average variable cost from given information on total cost.

Table 4.3 shows total cost at various outputs.

Table 4.3 Total costs of production for a firm

Output	Total cost (£)
0	100
10	115
20	140
30	175
40	220

From the information in the table, we can conclude that the average variable cost at an output of 20 is:

A £2
B £7
C £14
D £120

Only the third and first rows of data are used in the calculation — the data in the other rows are irrelevant. In row 3, we divide total cost (£140) by the output of 20, arriving at an average total cost of £7. The information in row 1 tells us that total fixed cost is £100. This means average fixed cost is £5 when 20 units of output are produced. Finally, since *AVC* = *ATC* – *AFC*, average variable cost is £7 – £5, which is £2. The correct answer is therefore A.

SECTION 4.4 SUMMARY

- Short-run cost curves are derived from short-run production theory.
- Long-run cost curves are derived from long-run production theory.
- Short-run costs are made up of fixed costs and variable costs.
- Fixed costs are the costs a firm incurs when hiring or paying for the fixed factors of production.
- Variable costs change as the level of output changes.
- Total costs comprise all the costs incurred when producing a particular level of output.
- Average cost is total cost divided by the level of output.
- Marginal cost is the addition to total cost resulting from producing one additional unit of output.
- The *ATC* and *AVC* curves are U-shaped with the *MC* curve cutting both at their lowest points.

4.5 Economies and diseconomies of scale

KEY TERMS

economies of scale as output increases, long-run average cost falls.

diseconomies of scale as output increases, long-run average cost rises.

As a firm grows in size by investing in new plant or buildings, it can benefit from economies of scale. However, beyond a certain size, the firm may eventually suffer from diseconomies of scale. **Economies of scale** are defined as *falling* long-run average costs of production that result from an increase in the size or scale of the firm. Likewise, **diseconomies of scale** are defined as *rising* long-run average costs of production.

Figure 4.10, which shows a 'symmetrical' U-shaped *LRAC* curve, has been drawn assuming that economies of scale are followed by diseconomies of scale. In the long run, as already explained, a firm can change the scale of all its factors of production, moving from one size of plant to another. At first sight, the *LRAC* curve appears identical to the short-run *ATC* curve drawn in Figure 4.9(a) and (b). However, the reasons for the U-shape of the short-run and the long-run

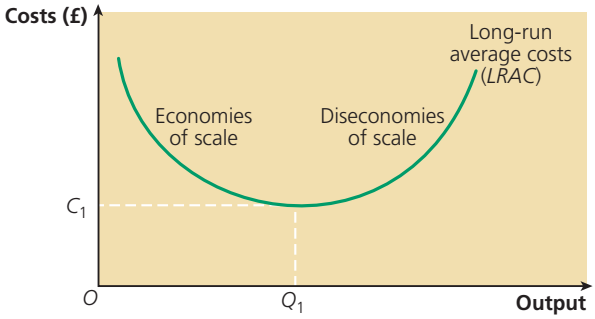


Figure 4.10 A firm's U-shaped long-run average cost (*LRAC*) curve

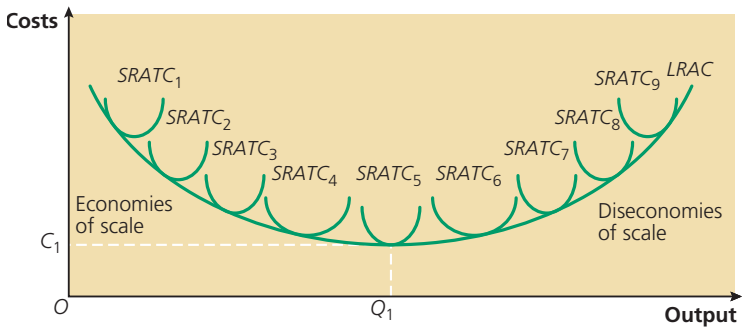


Figure 4.11 A firm's U-shaped long-run average cost (*LRAC*) curve and its related *SRATC* curves

average cost curves are different. The U-shape of the short-run *ATC* curve is explained by the assumption that labour becomes more productive as it is added to fixed capital, before eventually becoming less productive because of the impact of the law of diminishing returns. By contrast, the shape of the long-run average cost curve is explained by two long-run concepts which don't operate in the short run. These are economies and diseconomies of scale.

Figure 4.11 shows a number of short-run average total cost (*SRATC*) curves lying along the *LRAC* curve. Each of the *SRAS* curves, labelled $SRATC_1$ to $SRATC_9$, represents a particular firm size. In the long run, a firm can move from one short-run cost curve to another — for example, from $SRATC_1$ to $SRATC_2$ — with each curve associated with a different scale of capacity that is fixed in the short run. The *LRAC* curve forms a tangent to the *SRATC* curves, each of which touches the *LRAC* curve.

The difference between internal and external economies of scale

The scale economies and diseconomies referred to so far in this chapter have been **internal economies and diseconomies of scale**. To recap, these occur when a firm, or a plant within the firm itself, increases its scale and size. By contrast, **external economies of scale** occur when average or unit costs of production fall, not because of the growth of the firm or plant itself, but because of the growth of the industry or market of which the firm is a part. Likewise, **external diseconomies of scale** occur when average costs of production increase because of the growth of the whole industry or market. The next two sections explain different types of economy and diseconomy of scale.

KEY TERMS

internal economies and diseconomies of scale changes in long-run average costs of production resulting from changes in the size or scale of a firm or plant.

external economy of scale a fall in long-run average costs of production resulting from the growth of the market or industry of which the firm is a part.

external diseconomy of scale an increase in long-run average costs of production resulting from the growth of the market or industry of which the firm is a part.

Internal economies of scale

There are several types of internal economy of scale, as described below.

Technical economies of scale

Technical economies of scale are generated through changes to the 'productive process' as the scale of production and the level of output increase. Technical economies of scale can be caused by the following:

- **Indivisibilities.** Many types of plant or machinery are indivisible in the sense that there is a certain minimum size below which they cannot efficiently operate.
- **The spreading of research and development costs.** With large plants, research and development (R&D) costs can be spread over a much longer production run, reducing unit costs in the long run.
- **Volume economies.** With many types of capital equipment, costs increase less rapidly than capacity. When a storage tank or boiler is doubled in each of its three dimensions, its storage capacity increases eightfold. Volume economies are important in industries such as transport, storage and warehousing, as well as in metal and chemical industries where an increase in the scale of plant provides scope for the conservation of heat and energy.
- **Economies of massed resources.** The operation of a number of identical machines in a large plant means that proportionately fewer spare parts need to be kept than when fewer machines are involved.
- **Economies of vertically linked processes.** Much manufacturing activity involves a large number of vertically related tasks and processes, from the initial purchase of raw materials, components and energy through to the completion and sale of the finished product. The linking of processes in a single plant can lead to a saving in time, transport costs and energy.

Managerial economies of scale

The larger the scale of a firm, the greater is its ability to benefit from specialisation and the division of labour within management as well as within the ordinary labour force. A large firm can benefit from a functional division of labour, namely the employment of specialist managers: for example, in the fields of production, personnel and sales. Detail can be delegated to junior managers and supervisors.

Marketing economies of scale

Marketing economies of scale are of two types: bulk-buying and bulk-marketing economies. Large firms may be able to use their market power both to buy supplies at lower prices and also to market their products on better terms negotiated with wholesalers and retailers.

Financial or capital-raising economies of scale

Financial or capital-raising economies of scale are similar to the bulk-buying economies just described, except that they relate to the 'bulk-buying' or bulk borrowing of funds required to finance the business's expansion. Large firms can often borrow from banks and other financial institutions at a lower rate of interest and on better terms than those available to small firms.

Risk-bearing economies of scale

Large firms are usually less exposed to risk than small firms because risks can be grouped and spread. Large firms can spread risks by diversifying their

output, their markets, their sources of supply and finance, and the processes by which they manufacture their output. Such economies of diversification or risk bearing can make the firm less vulnerable to sudden changes in demand or conditions of supply that might severely harm a smaller, less diversified business.

Economies of scope

Economies of scope are factors that make it cheaper to produce a range of products together than to produce each one of them on its own. An example is businesses sharing centralised functions, such as finance or marketing.

TEST YOURSELF 4.4
Explain the difference between technical and managerial economies of scale.

Internal diseconomies of scale

Firms can also suffer from various types or forms of diseconomy of scale, or rising long-run average costs, as they grow in size. These include managerial diseconomies of scale, communication failure, and motivational diseconomies of scale.

Managerial diseconomies of scale

As a firm grows in size, administration of the firm becomes more difficult. Delegation of some of the managerial functions to people lower in the organisation may mean that personnel who lack appropriate experience make bad decisions. This may increase average costs of production.

Communication failure

Communication failure also contributes to managerial diseconomies of scale. In a large organisation, there may be too many layers of management between the top managers and ordinary production workers, and staff can feel remote and unappreciated. When staff productivity begins to fall, unit costs begin to rise. As a result, the problems facing the business are not effectively addressed.

Motivational diseconomies of scale

With large firms, it is often difficult to satisfy and motivate workers. Over-specialisation may lead to de-skilling and to a situation in which workers perform repetitive tasks and have little incentive to use personal initiative in ways which help their employer.



It can be difficult to motivate workers who perform repetitive, boring tasks

EXTENSION MATERIAL

Plant-level economies of scale and firm-level economies of scale

So far, we have discussed economies and diseconomies of scale which occur when the whole of a *firm* grows in size. Sometimes, however, *firms* grow larger but without the *plants* they own and operate growing significantly in size. For this reason, it is useful to distinguish between economies of scale that occur at the level of a single plant or establishment owned by a firm and those occurring at the level of the whole firm. In recent years, continued opportunities for further firm-level economies of scale have contributed to the growth of larger firms, but expansion of plant size has been less significant.

The relationship between returns to scale and economies or diseconomies of scale

The link between returns to scale and economies and diseconomies of scale is that increasing returns to scale lead to falling long-run average costs or economies of scale, and likewise decreasing returns to scale bring about rising long-run average costs or diseconomies of scale. The effect of increasing returns to scale on long-run average costs can be explained in the following way: output increases faster than inputs, so if wage rates and other factor prices are the same at all levels of output, the money cost of producing a unit of output must fall. Likewise, with decreasing returns to scale, output increases at a slower rate than inputs, and the money cost of producing a unit of output rises.

There are other reasons for falling long-run average costs besides the impact on costs of increasing returns to scale. These include the effect of 'bulk buying' reducing the cost of raw materials and components.

External economies and diseconomies of scale

As previously mentioned, external economies of scale occur when a firm's average or unit costs of production fall, not because of the growth of the firm itself, but because of the growth of the industry or market of which the firm is a part. Very often, external economies of scale are produced by cluster effects, which occur when a lot of firms in the same industry are located close to each other, providing markets, sources of supply and a pool of trained labour for each other.

External diseconomies of scale occur in a similar way, with the growth of the whole market raising the average costs of all the firms in the industry. As with external economies of scale, external diseconomies can arise from cluster effects. When a large number of similar firms locate close to each other, they not only create benefits which aid all the firms in the cluster; they may also get in each other's way. Competition for labour among the firms may raise local wages, which while being good for workers, increases the unit wage costs of their employers. There may also be an increase in local and regional traffic congestion, which lengthens delivery times and raises delivery costs both for firms and for their customers.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.4

Are economies of scale now less important in the car industry?

In the twentieth century, car manufacturing grew to become perhaps the most important industry in modern industrialised economies. Though car manufacture began in Germany and France, the main growth of car manufacturing in its early years took place in the USA. Henry Ford's adaptation of the **moving assembly line**, which allowed car factories to benefit from economies of scale, marked the beginning of **mass production**.

Garel Rhys, director of the Centre for Automotive Industry Research at Cardiff University, has calculated that economies of scale reach their peak at 250,000 cars a year in an assembly plant, although for the body panels the figure could be as high as 2 million.

However, economies of scale in car production are now not as important as they used to be. Reasons for this include: **market fragmentation**, leading to lower **production runs**; building cars to order rather than in large-scale batches of identical cars; and new ways of assembling finished cars in which manufacturers such as Toyota are outsourcing more and more of the car to outside suppliers. With car buyers demanding a wider choice of vehicles, production runs have to get smaller.

As car companies produce an ever-wider range of vehicles, so the way cars are made is changing. There is less need for huge, **capital-intensive factories**, and barriers to entry into the car industry are falling.

Follow-up questions

1. Explain the meaning of each of the concepts in bold type in the passage.
2. Describe some of the economies of scale that have contributed to lower average costs in the UK car industry.
3. Name three other manufacturing industries in which there are significant economies of scale and indicate in each case the types of economy of scale.
4. Outline two ways in which computers and robots have affected methods of production in the car industry.

Economies and diseconomies of scale and the long-run average cost curve

If you refer back to Figures 4.10 and 4.11, you will see symmetrical U-shaped *LRAC* curves, which have been drawn assuming that internal economies of scale are followed symmetrically by internal diseconomies of scale.

However, Figures 4.12 to 4.15 show four other possible shapes of the curve which differ from the symmetrical U-shape shown earlier. Figure 4.12 is a variant of Figures 4.10 and 4.11, but with a horizontal section to the *LRATC* curve inserted between the sections of the curve showing economies and diseconomies of scale. This depicts a situation between outputs Q_1 and Q_2 in which exhaustion of the benefits of economy of scale does not lead immediately to the onset of diseconomies of scale.

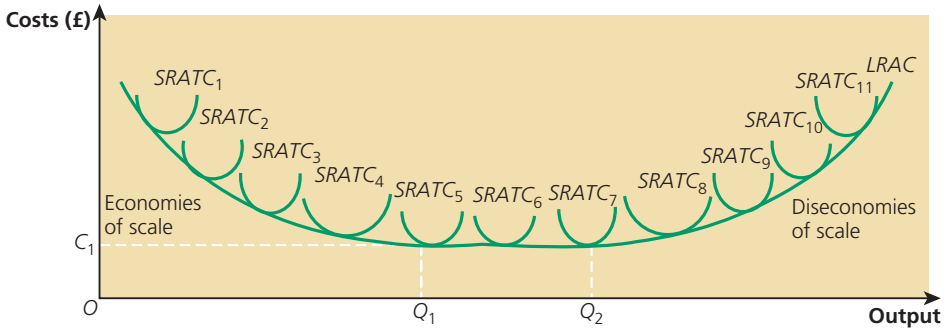


Figure 4.12 A 'three section' long-run average cost curve

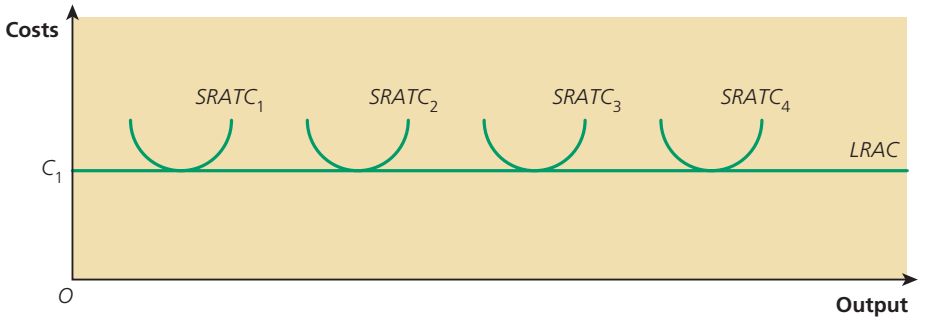


Figure 4.13 Constant long-run average costs

Another possibility is illustrated by Figure 4.13: an *LRAC* curve which is horizontal throughout its length. This curve depicts a market or industry in which firms neither benefit from economies of scale nor suffer the consequences of diseconomies of scale.

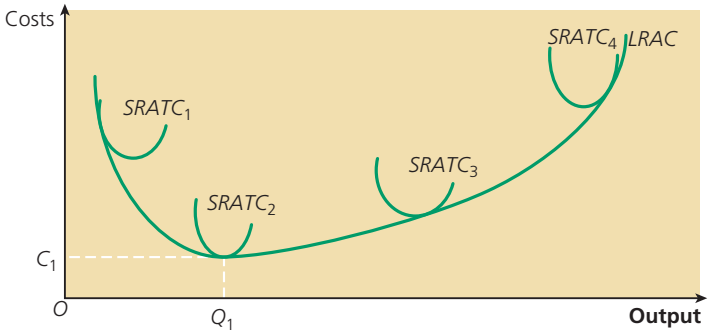


Figure 4.14 An *LRAC* curve showing economies of small-scale production

Figure 4.14 helps to explain why small firms are common in industries supplying services, such as those provided by hairdressers and personal trainers. The markets in which these services are provided typically possess economies of small-scale production. Diseconomies of scale set in early in such industries, leading to a U-shaped *LRAC* curve which is 'skewed to the left'.

By contrast, in Figure 4.15 the *LRAC* curve is 'skewed to the right' of the diagram, showing economies of large-scale production. Diseconomies of scale eventually set in, but only after substantial economies of scale have been achieved.

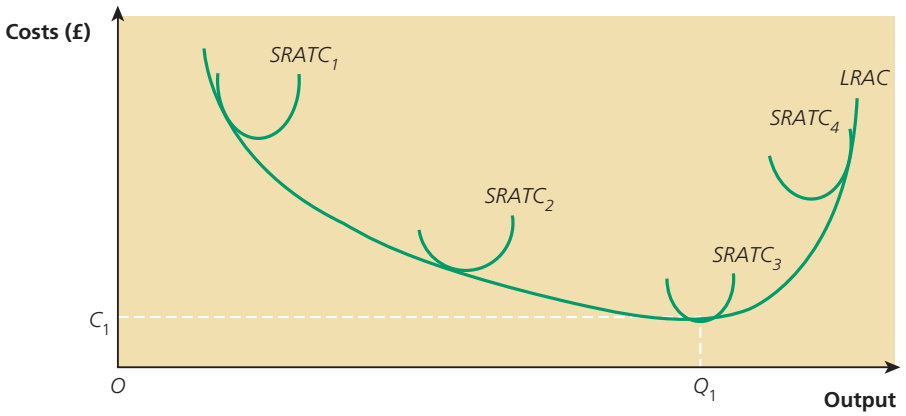


Figure 4.15 An *LRAC* curve showing economies of large-scale production

There are, of course, other factors, apart from the existence or non-existence of scale economies and diseconomies, which contribute to markets containing different sizes of firm. A factor that has been becoming increasingly important in recent decades is firms 'contracting out' the provision of services, previously provided 'in house' by managers and workers employed by the firms themselves, to 'outside' suppliers of the same services. The outside suppliers range from small independent firms such as a local sandwich shop to large-scale specialist firms providing services such as catering, accountancy and ICT maintenance.

All the *LRAC* curves drawn above can be used to show the *average cost-minimising* size of firm. In each case, this is shown at the lowest point on the

TEST YOURSELF 4.5
How many average cost-minimising sizes of firm are there in Figures 4.12 and 4.13? In each case, explain your answer.

LRAC curve and the *SRAC* curve forming a tangent to the *LRAC* curve at this point. In Figures, 4.10. 4.11, 4.14 and 4.15 there is only one *average cost-minimising* size of firm, located at output Q_1 with average cost C_1. Figures 4.12 and 4.13 are slightly different. Test Yourself 4.5 asks why this is so.

Long-run marginal costs

TEST YOURSELF 4.6
Apply the mathematical rule of the relationship between the marginal and average values of a variable (see page 88) to Figures 4.12–4.15. Copy each of the diagrams and then draw on the diagram the long-run marginal cost curve that fits the diagram.

Long-run marginal cost (*LRMC*) shows the additional cost incurred if a firm increases output when all the factors of production are variable. The mathematical relationship between *LRMC* and *LRAC* is the same as for *SRMC* and *SRATC*, and your understanding of this is tested in Test Yourself 4.6.

When long-run marginal costs fall and are *below* long-run average costs, the *LRAC* curve also falls. Likewise, when long-run marginal costs rise and are *above* long-run average costs, the *LRAC* curve also rises. It follows that when the *LRMC* curve cuts through the lowest point on a *LRAC* curve, the *LRAC* curve must be U-shaped, displaying first economies of scale and then diseconomies of scale.

How internal economies and diseconomies of scale affect particular industries

The following Applications of Economics in the Real World focuses on an industry that is affected by economies and diseconomies of scale.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.5

Super-tankers and volume economies of scale

The large super-tankers that are used to transport crude oil across seas and oceans from oil fields to industrial markets benefit significantly from volume economies of scale. However, super-tankers can also suffer from a diseconomy of scale. This is because large super-tankers cannot enter shallow ports. A wider tanker with a shallow draught does not yield as many economies of scale as a conventional super-tanker, but a wider tanker is more flexible and can enter more ports.

Follow-up questions

1. How could the oil tanker industry be organised to reduce the disadvantages of large super-tankers?
2. What is a volume economy of scale?
3. State three examples in other industries of volume economies of scale.
4. Explain how super-tankers sometimes cause *external* diseconomies of scale to occur.

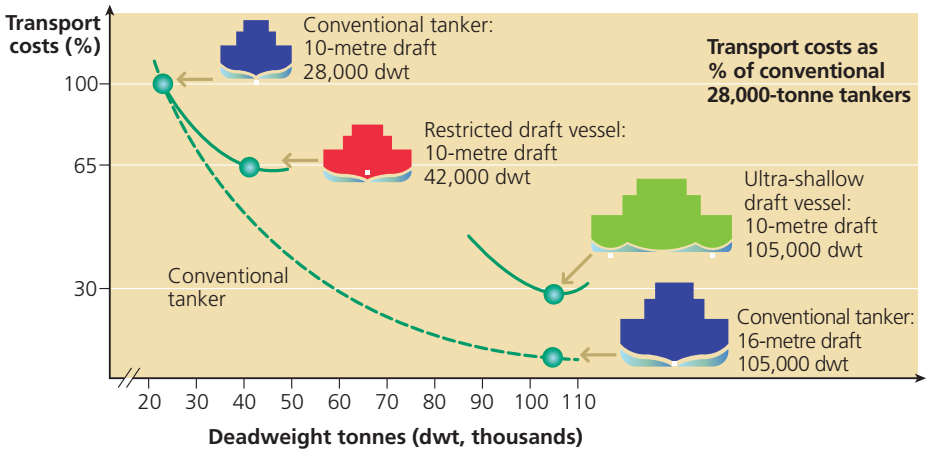


Figure 4.16 Economies of scale and oil tankers

The L-shaped long-run average cost curve

Besides the *LRAC* curves illustrated in Figures 4.10 to 4.15, another possibility is the L-shaped *LRAC* curve drawn in Figure 4.17. This curve, which vaguely resembles the letter 'L', results from the assumption that there are substantial economies of scale, which eventually give way, not to diseconomies of scale, but to a 'flattening out' of long-run average costs. There is evidence of this type of *LRAC* curve in many manufacturing industries involving large-scale production.

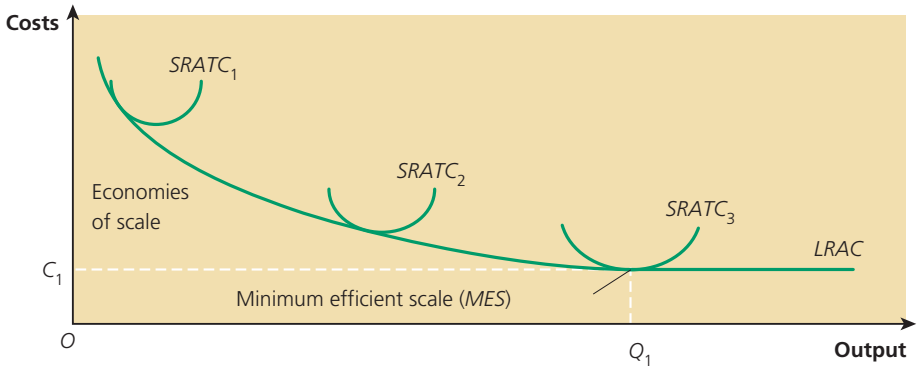


Figure 4.17 An 'L'-shaped *LRAC* curve and minimum efficient scale (*MES*)

The concept of the minimum efficient scale of production

Figure 4.17 also illustrates an important concept in production and cost theory: minimum efficient scale (*MES*). *MES* is the lowest output at which long-run average costs have been reduced to the minimum level that can be achieved, which means that the firm has benefited to the full from economies of scale. In Figure 4.17, all firm sizes to the left of the $SRATC_3$ curve are below minimum efficient scale, incurring higher average costs than can be achieved at the lowest point on $SRATC_3$. By contrast, there would be no further reductions in long-run production costs for any firms producing levels of output above Q_1. In the diagram, the *MES* level of output is Q_1, with average costs minimised at C_1.

SECTION 4.5 SUMMARY

- Economies of scale are associated with falling long-run average costs of production when a firm increases its size of plant.
- Diseconomies of scale are associated with increasing long-run average costs of production when a firm increases its size of plant.
- There are various types of economy of scale including technical, managerial and finance-raising economies.
- There are various types of diseconomy of scale including communicational failure and motivational diseconomies of scale.
- Economies and diseconomies of scale (in cost theory) are often respectively caused by increasing and decreasing returns to scale (in production theory).
- It is important not to confuse internal and external economies and diseconomies of scale.
- Likewise, it is important not to confuse short-run and long-run cost curves.
- In some industries, there are substantial economies of scale, which eventually give way to a 'flattening out' of long-run average costs. In these industries, there is a minimum efficient scale of production.

4.6 Marginal, average and total revenue

KEY TERMS

total revenue all the money received by a firm from selling its total output.

average revenue total revenue divided by output.

marginal revenue addition to total revenue resulting from the sale of one more unit of the product.

Revenue is the money that a firm earns when selling its output. **Total revenue** (*TR*) is all the money a firm earns from selling the total output of a product. By contrast, at any level of output, **average revenue** (*AR*) is calculated by dividing total revenue by the size of output. Stated as an equation:

$$\text{average revenue} = \frac{\text{total revenue}}{\text{output}} \text{ or } AR = \frac{TR}{Q}$$

Marginal revenue (*MR*) is the addition to total revenue resulting from the sale of one more unit of output. Stated as an equation:

$$\text{marginal revenue} = \frac{\Delta \text{ total revenue}}{\Delta \text{ output}} \text{ or } MR = \frac{\Delta TR}{\Delta Q}$$

where Δ is the symbol used to indicate the changes in total revenue and the change in total output.

STUDY TIP

The Greek delta symbol Δ is used by mathematicians as the symbol for a change in the value of a variable over a range of observations. The word 'marginal' means the change in the value of a variable when there is one more unit of the variable, so Δ is the symbol that indicates this change. It is used in the formulae for marginal product and marginal cost, as well as for marginal revenue.

Average revenue and the firm's demand curve

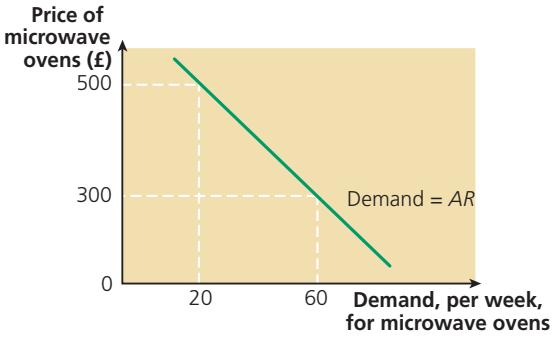


Figure 4.18 The demand curve facing a firm when it decides how much output to sell

Figure 4.18 shows the demand conditions facing a high-street retailer of electrical goods in a particular week. If the firm sets the price of microwave ovens at £500, 20 customers want to buy an oven. Total sales revenue is £10,000 and average revenue per oven sold is £500. However, by reducing the price to £300, the retailer can sell 60 ovens. Total revenue is now £18,000 and average revenue is £300. The point to note is that, at each level of sales, the average revenue the retailer earns is the same as the price charged. This is always the case when the price charged is the same for all the units of output being sold. Hence, in these conditions, the demand curve facing the firm is also its average revenue (*AR*) curve.

STUDY TIP

Make sure you understand that the demand curve for a firm's output is the same as the average revenue curve facing the firm.

The relationship between average and marginal revenue

To understand the relationship between marginal and average *revenue* curves you should start by rereading the explanation in the extension material on page 87 about the relationship between marginal and average *returns*. (Indeed, the relationship is the same for all marginal and average values of two variables, provided they are plotted from the same data set: for example, marginal and average cost plotted from cost data.)

Marginal and average curves plotted from the same set of data always display the following relationship:

- When the marginal is greater than the average, the average rises.
- When the marginal is less than the average, the average falls.
- When the marginal equals the average, the average is constant, neither rising nor falling.

The nature of a firm's revenue curves depends on the competitiveness of the market structure in which the firm sells its output. The bottom part of Figure 4.1 at the start of this chapter sets out the four market structures that we will explain in greater detail in Chapter 5. Two of these are **perfect competition** and **monopoly**. The marginal and average revenue curves are different in these market structures.

KEY TERMS

perfect competition a market that displays the six conditions of: a large number of buyers and sellers; perfect market information; the ability to buy or sell as much is desired at the ruling market price; the inability of an individual buyer or seller to influence the market price; a uniform or homogeneous product; and no barriers to entry or exit in the long run.

monopoly one firm only in a market.

Average revenue and marginal revenue in perfect competition

A perfectly competitive market is defined by a number of conditions or characteristics that the market must possess. These conditions, which we will revisit in Chapter 5, are:

- a very large number of buyers and sellers
- all buyers and sellers possess perfect information about what is going on in the market
- consumers can buy as much as they wish to purchase and firms can sell as much as they wish to supply at the ruling market price, set in the market as a whole
- an individual consumer or supplier cannot affect the ruling market price through its own actions
- an identical, uniform or homogeneous product
- no barriers to entry into, or exit from, the market in the long run.

Taken together, the six listed conditions tell us that a perfectly competitive firm, which is depicted in Figure 4.19(a), faces a perfectly elastic demand curve for its product. The demand curve facing the firm is located at the ruling market price, P_1, which itself is determined through the interaction of market demand and market supply in Figure 4.19(b). Note that the horizontal axis in Figure 4.19(b) shows millions of units of output being produced. This is because Figure 4.19(b) depicts the whole market, comprising very large numbers of both consumers and firms. In equilibrium, where market demand equals market supply, the ruling market price is P_1 and the equilibrium quantity is Q_1 millions of units. In Figure 4.19(a), the horizontal axis is labelled 'hundreds', to reflect the fact that in perfect competition a single firm is only a tiny part of the total market.

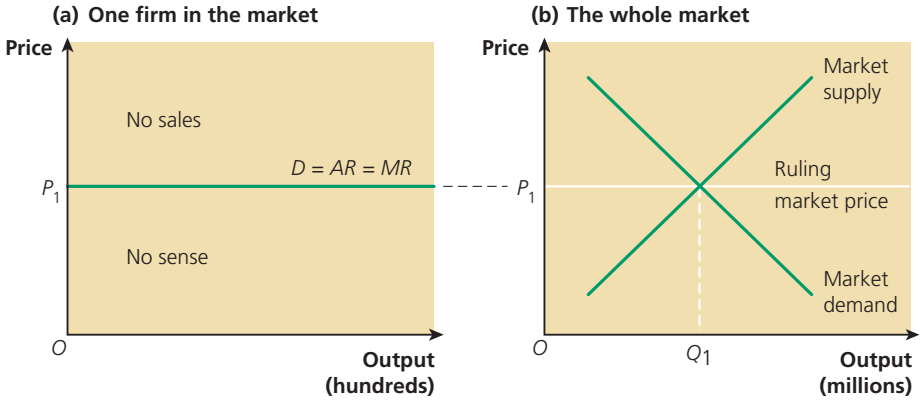


Figure 4.19 Average revenue and marginal revenue curves in perfect competition

KEY TERM
price-taker a firm which is so small that it has to accept the ruling market price. If the firm raises its price, it loses all its sales; if it cuts its price, it gains no advantage.

The assumption that a perfectly competitive firm can sell whatever quantity it wishes at the ruling market price P_1, but that it cannot influence the ruling market price by its own action, means that a firm in a perfectly competitive markets is a passive **price-taker**.

The labels 'No sales' and 'No sense' placed on Figure 4.19(a), respectively above and below the price line P_1, help to explain why a perfectly competitive firm is a price-taker. 'No sales' indicates that, if the firm raises its selling price above the ruling market price, customers desert the firm to buy the identical products (perfect substitutes) available from other firms at the ruling market price. 'No sense' refers to the fact that, although a perfectly competitive firm *could* sell its output below the price P_1, doing so is inconsistent with the profit-maximising objective. No extra sales can result, so selling below the ruling market price inevitably reduces both total sales revenue and therefore profit, given the fact that the firm can sell any quantity it wants at the ruling market price.

SYNOPTIC LINK
A perfectly competitive market for a good is explained in detail in Chapter 5, while Chapter 6 applies the concept of perfect competition to labour markets.

It follows from this that the ruling market price facing each firm in the market is both the firm's average revenue curve and its marginal revenue curve. If each unit of the good is sold at a price of £1 (average revenue), selling an extra unit of the good always increases total revenue by £1 (marginal revenue). The horizontal price line is also the perfectly elastic demand curve for the firm's output. It is perfectly elastic because the goods produced by all the firms in the market, being uniform or homogeneous, are perfect substitutes for each other. In summary, for a firm, $D = AR = MR$, as depicted in Figure 4.19(a).

STUDY TIP
Note that Figure 4.19 contains two demand curves. On the one hand, the market demand curve, drawn in Figure 4.19(b), slopes downwards, reflecting the fact that in the market as a whole a fall in price will lead to an increase in the quantity sold. With a straight-line market demand curve, the price elasticity of demand falls from point to point moving down the demand curve. Goods produced in other markets are partial, though not perfect, substitutes for goods produced in this market. By contrast, the demand curve for the output of any one firm *within* the market, shown in Figure 4.19(a), is perfectly elastic, and hence horizontal, because other firms *within* the market produce identical goods which are perfect substitutes for the goods produced by any one firm in this market.

Average revenue and marginal revenue in monopoly

It is worth repeating that the demand curve facing a perfectly competitive firm, besides being located at the ruling market price, is also the firm's *AR* curve and

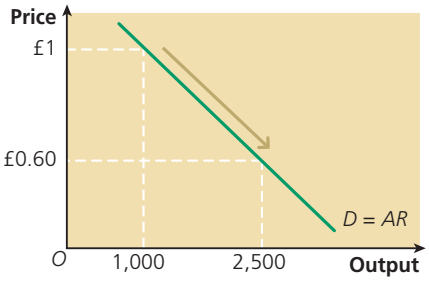


Figure 4.20 Price equals average revenue (*AR*) in monopoly

its *MR* curve. By contrast, the demand curve for a monopolist's output is the monopolist's *AR* curve, but it is *not* the monopolist's *MR* curve.

To understand why the market demand curve is the monopolist's average revenue (*AR*) curve, consider Figure 4.20, which shows two prices, £1 and £0.60, which can be charged by a monopolist for the good it produces.

At a price of £1, 1,000 units are demanded. At this price, the monopolist's total revenue is £1,000. Average revenue, or total revenue divided by output ($TR \div Q$), is £1, which is of course the same as price. This is the case at all prices: the price charged for all units of the good and average revenue are always the same. For example, if the monopolist sets the price at £0.60, 2,500 units of the good are demanded; total sales revenue is £1,500 and average revenue ($TR \div Q$) is £0.60. The downward-sloping market demand curve facing the monopolist is therefore the firm's average revenue (*AR*) curve.

KEY TERMS

price-maker when a firm faces a downward-sloping demand curve for its product, it possesses the market power to set the price at which it sells the product.

quantity-setter when a firm faces a downward-sloping demand curve for its product, it possesses the market power to set the quantity of the good it wishes to sell.

The downward-sloping *AR* curve can affect the monopoly in two different ways. If the monopolist is a **price-maker**, choosing to set the price at which the product is sold, the demand curve dictates the maximum output that can be sold at this price. For example, if the price is set at P_1 in Figure 4.21, the maximum quantity that can be sold at this price is Q_1. But if the monopolist cuts the price it charges to P_2, sales increase to Q_2. Alternatively, if the monopolist is a **quantity-setter** rather than a price-maker, the demand curve dictates the maximum price at which a chosen quantity of the good can be sold. If the monopolist wants to sell Q_2, the market demand curve shows that the maximum price at which this quantity can be sold is P_2. To summarise, if the monopolist sets the price, the market demand curve dictates the maximum quantity the firm can sell. Conversely, if the monopolist sets the quantity, the market demand curve determines the maximum price the firm can charge. However, for any one good it produces, a firm cannot be a price-maker and a quantity-setter at the same time.

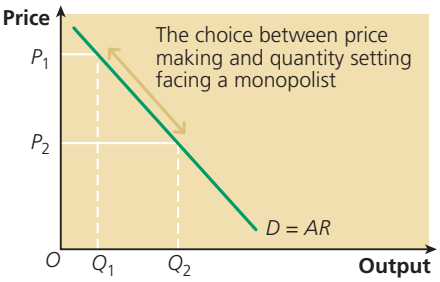


Figure 4.21 The choice between price making and quantity setting facing a monopolist

However, to understand why marginal revenue and average revenue are *not* the same in monopoly, you must remember that, when the *marginal* value of a variable is less than the *average* value of the variable, the *average* value falls.

Because the market demand curve or average revenue curve falls as output increases, the monopolist's marginal revenue curve *must* be below its average revenue curve. Figure 4.22 shows the relationship between the *AR* and the *MR* curves. You should see, however, that the *MR* curve is not only below the *AR* curve — it has also been drawn twice as steep. This is always the case whenever the *AR* curve is a downward-sloping straight line.

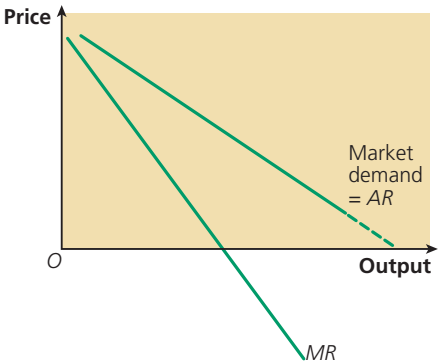


Figure 4.22 Monopoly average revenue (*AR*) and marginal revenue (*MR*) curves

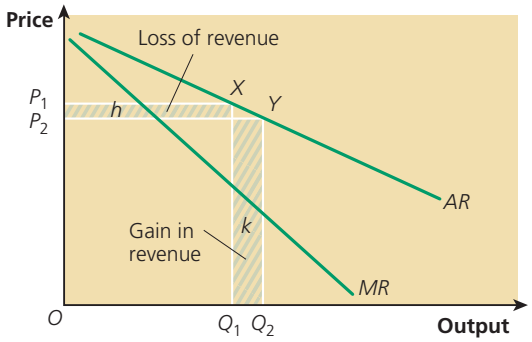


Figure 4.23 Explaining a monopolist's marginal revenue (*MR*) curve

Figure 4.23 explains in more depth the relationship between *AR* and *MR* curves in monopoly. The monopolist initially charges a price of P_1 and sells the level of output Q_1. However, to increase sales by an extra unit to Q_2, the downward-sloping *AR* curve forces the monopolist to reduce the selling price to P_2. This reduces the price at which *all* units of output are sold. Total sales revenue increases by the area *k* in Figure 4.23, but decreases by the area *h*. Areas *k* and *h* respectively show the revenue gain (namely, the extra unit sold multiplied by price P_2) and the revenue loss resulting from the fact that, in order to sell more, the price has to be reduced for *all* units of output, not just the extra unit sold. Marginal revenue, which is the revenue gain minus the revenue loss (or *k* – *h*), must be less than price or average revenue (area *k*).

SYNOPTIC LINK

A monopoly is explained in detail in Chapter 5, while Chapter 6 applies the related concept of monopsony to labour markets.

EXTENSION MATERIAL

Elasticity and revenue curves

We mentioned earlier in the chapter that the horizontal price line facing a perfectly competitive firm is also the perfectly elastic demand curve for the firm's output. The explanation for this lies in the word 'substitutability'. When studying elasticity, you learnt that the availability of substitutes is the main determinant of price elasticity of demand. In perfect competition, because of the assumptions of a uniform product and perfect information, the output of every other firm in the market is a *perfect substitute* for the firm's own product. If the firm tries to raise its price above the ruling market price, it loses all its customers.

In monopoly, by contrast, provided the demand curve is a straight line as well as downward sloping, price elasticity of demand falls moving down the demand curve. Demand for the monopolist's output is elastic in the top half of the curve, falling to unit elastic exactly half way down the curve, and inelastic in the bottom half of the curve. This is shown in Figure 4.24. Demand is elastic between *A* and *B*, unit elastic at *B*, and inelastic between *B* and *C*.

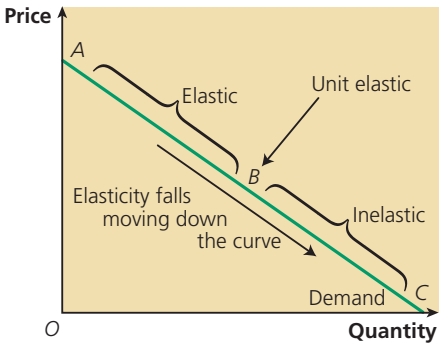


Figure 4.24 Price elasticity of demand and a monopolist's demand or average revenue (*AR*) curve

SYNOPTIC LINK

We will revisit the significance of elasticity in the next chapter, when comparing profit maximisation with revenue maximisation.

The relationship between marginal revenue and total revenue

Marginal revenue measures the change in total revenue that results from an increase in the quantity of goods sold. It indicates how much revenue increases for selling an additional unit of a good or service. Marginal revenue is also shown by the *slope* of the total revenue curve. Increasing marginal revenue can be shown by the total revenue curve becoming steeper. Falling marginal revenue is shown by the total revenue curve becoming less steep as sales increase. And finally, constant marginal revenue means that the slope of the total revenue curve is unchanged as sales increase. Look back to Figure 4.4 on page 86 of this chapter. This illustrates the relationship between marginal returns and total returns — another example of the relationship between the marginal and the total values of a variable. The relationship between marginal costs and total cost provides a third example.

STUDY TIP

Make sure you practise calculating and plotting average, marginal and total revenue from given data. This will help you to build up an appreciation of how they are interrelated.

SECTION 4.6 SUMMARY

- Revenue is the money that a firm earns when selling its output.
- Total revenue is all the money a firm earns from selling the total output of a product.
- Average revenue is calculated by dividing total revenue by the level of output.
- Marginal revenue is the addition to total revenue resulting from the sale of one more unit of the product.
- The demand curve facing the firm is also its average revenue curve.
- The nature of a firm's revenue curves depends on the competitiveness of the market structure in which the firm sells its output.
- In perfect competition, the average and marginal revenue curves are horizontal.
- In monopoly, the average revenue curve slopes downwards, with the marginal revenue curve below the average revenue curve.

4.7 Profit

Profit is the difference between total revenue and total costs

Students often confuse profit and revenue, mistakenly believing that the two terms have the same meaning. **Profit** is the difference between the sales revenue the firm receives when selling the goods or services it produces and the costs it incurs when producing these goods or services.

total profit = total revenue – total costs

In some circumstances, total costs of production may exceed total sales revenue, in which case there is a loss. Think of a loss as being *negative* profit.

We have already mentioned that economists generally assume that firms have a single business objective: **profit maximisation**. This means producing the

KEY TERMS

profit the difference between total sales revenue and total cost of production.

profit maximisation occurs at the level of output at which total profit is greatest.

level of output at which profit (revenue minus costs) is greatest. (Firms may also have other objectives, such as survival, growth and increasing their market share.)

The difference between normal and abnormal (supernormal) profit

When explaining profit maximisation, we will apply two profit concepts, used frequently by economists undertaking microeconomic analysis, but rarely used outside the field of microeconomic theory. These are *normal profit* and *abnormal profit*.

Normal profit is the minimum level of profit necessary to keep established firms in the market, rewarding the time, decision making and entrepreneurial risk taking 'invested' into production. However, the normal profit made by firms already established in the market is insufficient to attract new firms into the market. Economists treat normal profit as an opportunity cost, which they include in firms' average cost curves. In the long run, firms unable to make normal profit leave the market. Normal profit varies from one industry to another, depending on the risks facing firms.

Abnormal profit, or **supernormal profit**, by contrast, is extra profit over and above normal profit.

KEY TERMS

normal profit the minimum profit a firm must make to stay in business, which is, however, insufficient to attract new firms into the market.

abnormal profit (also known as **supernormal profit** and **above-normal profit**) profit over and above normal profit.

STUDY TIP

Avoid confusing normal profit with another abstract microeconomic term: normal good. You came across normal goods when studying demand theory in Chapter 3. Normal goods are types of goods which experience an increase in demand as consumers' incomes increase.

QUANTITATIVE SKILLS 4.5

Worked example: calculating revenue and profit

Table 4.4 provides information about the short-run output, costs and revenue of a firm.

Table 4.4 Short-run output, costs and revenue

Output per week	Total revenue (£000s)	Total cost (£000s)
0	0	10
1	20	14
2	38	19
3	54	28
4	68	44
5	80	80
6	90	93

From the information in the table, calculate:

(a) marginal revenue when output per week increases from 4 to 5 units

The marginal revenue, which is the change in the total revenue, is £80,000 – £68,000, which is £12,000.

(b) the level of output at which the firm would make normal profit but not abnormal profit

Assuming that normal profit is being treated as a cost of production, the firm makes normal profit, but not abnormal profit, when total revenue equals total cost. This is at a level of output of 5 units per week.

(c) the profit-maximising level of output per week

Profits are maximised at the level of output at which (*TR* – *TC*) is greatest. This is an output of 3 units per week, when total profit equals £54,000 – £28,000, which is £26,000.

The role of profit in a market economy

In the long run, and in the absence of entry barriers, abnormal profit performs the important economic function of attracting new firms into the market. Not only do rising profits, and the hope of higher profits in the future, provide business incentives for managers within a firm to work harder to make the business even more profitable, they also create incentives for other firms to enter the market. Abnormal profit acts as a 'magnet' attracting new entrants into a market or industry. As Chapter 5 explains, if market entry is easy and/or relatively costless, new firms joining the market should lead to an increase in market supply. In competitive markets, the entry of new firms triggers a process which reduces both abnormal profit and prices, with the latter benefiting consumers.

However, when entry barriers are high and monopoly or highly imperfect competition exists in a market, profit may simply reward inefficient producers. This is a form of market failure in which the producer rules the market rather than the consumer, and in which 'producer sovereignty' rather than 'consumer sovereignty' exists.

Except when monopolies make large profits by exploiting their consumers, profit can be an indicator of economic efficiency. Large profits might mean that firms have succeeded in eliminating unnecessary costs of production and are also using the most efficient production processes.

Profit performs other roles in a market economy. These include the creation of worker and shareholder incentives. Profit also influences the allocation of resources, it is an efficiency indicator, and it is a reward for innovation and for risk taking. Finally, profits provide an important source of business finance.

The creation of worker incentives

Some companies use profit-related pay and performance-related pay to increase worker motivation, in the hope that workers will work harder and share the objectives of the business's managers and owners. This can, however, be counterproductive if ordinary workers see higher management and company directors enjoying huge profit-related bonuses, while they receive low wages.

The creation of shareholder incentives

High profit generally leads to high dividends or distributed profit being paid out to shareholders who own companies. This creates an incentive for more people to want to buy the company's shares. As a result, the company's share price rises, which makes it cheaper and easier for a business to raise finance.

Profits and resource allocation

High profits made by incumbent firms in a market create incentives for new producers to enter the market and for existing firms to supply more of a good or service. Likewise, loss making, or perhaps a failure to make abnormal profits, creates incentives for firms to leave markets and to deploy their resources in more profitable markets.

Profit as a reward for innovation and risk taking

As we explain in the final section of this chapter, innovation is an improvement on something that has already been invented, which thus turns the results of invention into a useful product. If entrepreneurs believe that innovation can result in high profits in the future, the incentive to innovate

TEST YOURSELF 4.7
How may increases in economic efficiency and economic welfare be promoted by the pursuit of profit?

increases. As we can never be sure of future profits, risks are involved. However, successful risk taking leads to high profits.

Profit as a source of business finance

Instead of being distributed to the business's owners as a form of income, profit can be retained within the business. Retained profits are perhaps the most important source of finance for firms undertaking investment projects. High profits also make it easier and cheaper for firms to use borrowed funds as an important source of business finance.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.6

The John Lewis economy

John Lewis PLC is a generally successful retailing company which shares its profits with all its workers and makes them part-owners, within the business, of the John Lewis Partnership. The letters PLC stand for public limited company, which means the company's shares can usually be bought and sold on the stock market. But John Lewis does not have external shareholders. Instead the Partnership offers its employees indirect ownership of the company. They do not directly own shares in the business, but are granted some voting rights at board level plus some of the profits, which are added to their salaries each year.

The John Lewis Partnership is the largest employee-owned business in the UK. Government ministers have claimed that employee share ownership is key to improving company profits and productivity. But as it stands, the employee-owned sector is not currently large enough to bring about a major change in the character of the British business model.

In recent years John Lewis has been less successful than it was a few years ago. Profits have fallen, so worker pay has also fallen. Plans for a 'John Lewis economy' that would inspire businesses to share ownership with their staff have also been undermined by the removal of government support services offered to such businesses.



John Lewis offers its employees part-ownership of the business

Follow-up questions

1 Explain the term 'profit sharing'.
2 Outline one advantage and one disadvantage of profit sharing and co-ownership for the UK economy.
3 Find out more details of the economic performance of John Lewis PLC in recent years.
4 Why is John Lewis generally considered to be an unusual form of PLC?

SECTION 4.7 SUMMARY

- Profit is not the same as revenue.
- Profit is revenue minus cost.
- In traditional economic theory, but not necessarily in behavioural theory, firms are assumed to be profit maximisers.
- Normal profit is the minimum profit a firm must make to stay in business, which is, however, insufficient to attract new firms into the market.
- Abnormal profit is profit over and above normal profit.
- Profit has various roles in a market economy, including creating incentives and rewarding innovation and risk taking.

4.8 Technological change

KEY TERM
technological change a term used to describe the overall effect of invention, innovation and the diffusion or spread of technology in the economy.

Most people have a general idea of what technology means, but they nevertheless find it difficult to give the term a precise definition. This is one definition: whereas science is concerned with how and why things happen, technology focuses on *making things happen*. Thus, technology is knowledge put to practical use to solve problems facing human societies.

Technological change, by contrast, involves changing existing technologies, hopefully for the better, and the development of completely new technologies, both to improve existing *products* and the *processes* involved in making the products, and to develop completely new products and processes. In the economic sphere, this leads to the development of completely new markets, to changes in market structure, and also to the destruction of some existing markets.

Technological change and technical progress

The word 'technology' is often associated with technical progress. However, the term 'technical progress' has two rather different meanings. On the one hand, in a value-judgement context, 'technical progress' implies that technological change is fundamentally about increasing economic welfare and making people happier. For example, although the development and use of the motor car has several important drawbacks, such as the harm resulting from road accidents and environmental pollution, for the most part, through making it much easier for most people to travel, cars and buses have significantly improved human welfare.

But in a narrower sense, unrelated to welfare considerations, 'technical progress' means applying scientific and engineering knowledge, as it develops, to produce goods which are more efficient and work better, regardless of whether these are good for society. In this narrower meaning, technical progress includes the development of distinctly harmful goods such as chemical weapons, which, when used, have a devastating effect on human welfare.

The difference between invention and innovation

KEY TERMS
invention making something entirely new; something that did not exist before at all.
innovation improves on or makes a significant contribution to something that has already been invented, thereby turning the results of invention into a product.

Invention is about creating new ideas for products or processes. **Innovation**, by contrast, converts the results of invention into marketable products or services. Many inventions fail to see the light of day because they have no practical use. It is innovation, rather than invention, that is really important for the success of a firm.

In an article in *Huffpost* published on 6 December 2017, the American entrepreneur Tom Grasty distinguishes between invention and innovation in the following way:

> In its purest sense, invention can be defined as the creation of a product or introduction of a process for the first time. Innovation, on the other hand, occurs if someone improves on or makes a significant contribution to an existing product, process or service. Consider

the microprocessor. Someone invented the microprocessor. But by itself, the microprocessor was nothing more than another piece on the circuit board. It's what was done with that piece — the hundreds of thousands of products, processes and services that evolved from the invention of the microprocessor — that required innovation.

If ever there were a poster child for innovation it would be former Apple CEO Steve Jobs. And when people talk about innovation, Jobs' iPod is cited as an example of innovation at its best. But let's take a step back for a minute. The iPod wasn't the first portable music device (Sony popularized the 'music anywhere, anytime' concept 22 years earlier with the Walkman); the iPod wasn't the first device that put hundreds of songs in your pocket (dozens of manufacturers had MP3 devices on the market when the iPod was released in 2001); and Apple was actually late to the party when it came to providing an online music-sharing platform (Napster, Grokster and Kazaa all preceded iTunes).

So, given those sobering facts, is the iPod's distinction as a defining example of innovation warranted? Absolutely. What made the iPod and the music ecosystem it engendered innovative wasn't that it was the first portable music device. It wasn't that it was the first MP3 player. And it wasn't that it was the first company to make thousands of songs immediately available to millions of users. What made Apple innovative was that it combined all of these elements — design, ergonomics and ease of use — in a single device, and then tied it directly into a platform that effortlessly kept that device updated with music.

Apple invented nothing. Its innovation was creating an easy-to-use ecosystem that unified music discovery, delivery and device. And, in the process, they revolutionized the music industry.

The effect of technological change on economic performance

Through its diffusion into the economy, technological change affects methods of production, productivity, efficiency and firms' costs of production. We will now look at each of these in turn.

Methods of production

Throughout human history, technological change has affected methods of production. The mid-twentieth century witnessed the growing use of cars. We are now living in the 'computer age'. Computers, which were first developed in the 1940s, are now widely used in manufacturing (e.g. when computer-controlled robots build cars), in distribution (e.g. in the online sale of books by Amazon), and as consumer goods in themselves. And hidden within many goods that are not themselves computers, such as washing machines and cars, are microprocessors that control how the good functions.

Closely allied to the changes in production has been the change in recent decades from mechanised to automated production. As a simplification, mechanisation means that human beings operate the machines that are used to produce goods. Automation, by contrast, means that machines operate other machines — for example, a computer-controlled robot operating a welding tool to weld together the body panels of a car. Both mechanisation and automation have often been accompanied by assembly-line production, allegedly first introduced by Henry Ford in 1908.

STUDY TIP

Make sure you don't confuse *production* with *productivity*. The two concepts are closely related, but production refers to total output, whereas productivity is output per unit of input.

Productivity

Earlier in this chapter on page 78, we reminded you of the meaning of *productivity*. We defined productivity as output per unit of input, though it can also be considered as output per unit of time. We also said that, when economists talk about productivity, they usually mean *labour productivity*. Technological change generally increases labour productivity. This has usually been the case following the introduction of both mechanised and automated production methods. However, in the case of automation and the use of computers in production, there have been several well-publicised examples in organisations such as the National Health Service of very expensive computer systems that have failed to work properly and which, in extreme cases, have had to be scrapped. In these cases, labour productivity may fall rather than increase, at least until the system can be made to work properly.

Efficiency

KEY TERMS

productive efficiency for the economy as a whole occurs when it is impossible to produce more of one good without producing less of another. For a firm it occurs when the average total cost of production is minimised.

dynamic efficiency measures improvements in productive efficiency that occur in the long run over time.

As we explain in Chapter 5, economists recognise a number of types of economic efficiency. Two of these are productive efficiency and dynamic efficiency. **Productive efficiency** centres on minimising average costs of production. **Dynamic efficiency** measures the extent to which productive efficiency increases over time, in the economic long run. Dynamic efficiency results from improvements in products and services, innovation and the process of creative destruction.

Technological change generally improves both productive efficiency and dynamic efficiency. As a general rule, technological change leads to improvements in both productive and dynamic efficiency. By increasing productivity, over time technological changes shift both short-run and long-run cost curves downwards, thereby improving both productive and dynamic efficiency.

Costs of production

It follows from the fact that technological change generally improves both productivity and efficiency that it also reduces costs of production, in the short run but especially in the long run. This is because the long run is the time period in which firms can invest in new capital equipment and technology which they hope will reduce costs of production.

EXTENSION MATERIAL

Some other effects of technological change

A theme running through this section on technological change is that, particularly in recent years, technological change has been highly significant in the development of new products and new markets, and the destruction of existing markets. To explain this further, it is useful to introduce the concepts of *disruptive* innovation and *sustaining* innovation.

A disruptive innovation is an innovation that helps create a new market, but in so doing eventually disrupts an existing market over a few years or decades, thereby displacing an earlier technology. Disruptive innovation often improves a product or service in ways that the market did not initially expect. It creates new goods or services for a different set of consumers in a new market which competes with the established market. By doing so, it eventually lowers prices in the existing market. By contrast, a sustaining innovation does not create new markets but develops existing markets, enabling firms within them to offer better value and often to compete against each other, sustaining improvements.

According to Harvard University business professors Joseph L. Bower and Clayton M. Christensen, one of the most consistent patterns in business is the

failure of leading companies to stay at the top of their industries when technologies or markets change. Writing back in the 1990s, Bower and Christensen gave the example of the US company, Xerox, which at the time had dominated the photocopier market, losing market share to the Japanese company, Canon, in the small photocopier market.

Bower and Christensen ask why it is that companies like Xerox invest aggressively — and successfully — in the technologies necessary to retain their current customers, but then fail to make certain other technological investments that customers of the future will demand. The explanation they offer is that companies that dominate an existing technology are in danger, as disruptive innovation occurs, of remaining too close to their existing body of customers. All too often existing customers reject the goods produced by a new technology because it does not address their needs as effectively as a company's current products. The large photocopying centres that represented the core of Xerox's customer base at first had no use for small, slow table-top copiers produced by Xerox's new technology. Result: Canon stepped in, quickened the speed of the copiers, and took over the market.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.7

The impact of the smartphone on Kodak and the traditional camera industry

In September 2013, the American camera company Kodak emerged from the bankruptcy it had been in for nearly two years. Since 2000, demand for Kodak's most successful product, camera film, which once ranked among the most profitable consumer products ever invented, had been in rapid decline. To make matters worse, Kodak's management had grossly under-estimated the speed of the collapse.

This was all due to the development of digital cameras, and later smartphones. Global sales of traditional photographic film and paper dropped like a stone. Up to that point, the boxes of film that Kodak produced had been highly profitable. New entrants were deterred by the high costs of entry to this capital-intensive market and Kodak enjoyed profit margins that were sometimes as high as 50%.

But in the first few months of this century, a technology change began that was to wreck one of the most lucrative business models of the last century and threaten Kodak's very existence. In 1999, only 5% of new cameras sold in the USA were digital. By the end of 2000, this had changed dramatically. By 2003, now being accused of ignoring the revolution until it was too late, Kodak cut tens of thousands of jobs at its capital-intensive film factories and announced a new digital strategy.

In 2016, Kodak sold its first smartphone, the Ektra, with a design based on the iconic 1941 Kodak Ektra 35mm camera. Kodak was aiming to establish a nostalgic niche in a highly competitive market dominated by Apple's iPhone and Samsung smartphones. Kodak said it wanted to make the Ektra for people with 'artistic-oriented hobbies, interests, passions'. However, Kodak's Ektra smartphone was badly reviewed, and despite significant price cuts and upgrades, sales were poor. So far, Kodak has failed to establish its niche in the smartphone market.

Disruptive digital technology has caused the crash of many business models since 1999 — but few quite so rapid as the fate that befell Kodak. Within a matter of months, the once hugely profitable camera film market had given way to the surge of digital cameras and smartphones.



The Ektra, Kodak's entry into the smartphone market

Follow-up questions

1. Find out fuller details of what has happened to Kodak since the company emerged from bankruptcy in 2013.
2. In what way have smartphones, which were first marketed in 2007, affected the market for digital cameras?
3. Research statistics for camera sales in a recent year by visiting **www.statista.com**. Construct a pie chart to show the sales of different camera manufacturers.
4. Similarly, research statistics for smartphone sales in a recent year. Construct a pie chart to show the sales of different smartphone manufacturers.

The influence of technological change on the structure of markets

KEY TERMS

monopolistic competition a market structure in which firms have many competitors, but each one sells a slightly different product.

duopoly two firms only in a market.

Application of Economics in the Real World 4.7 on Kodak provides a good example of the influence of technological change on the structure of markets. It describes how, when cameras transmitted the images they photographed onto chemical film, very high entry barriers into the chemical film market led to Kodak's domination of the market. By contrast, entry into the digital camera market is relatively easy. Hence the camera film market, dominated by Kodak, was close to a monopoly, whereas the digital camera and smartphone markets are closer to a much more competitive form of market, **monopolistic competition**, which we explain in Chapter 5.

Technological change does not always, however, lead to more competitive market structures. In some industries, technological change has led to outcomes in which very large firms dominate. This happens if technological change leads to capital indivisibilities, which occur when very large quantities of capital equipment are required for one unit of a good to be produced. It is not economically feasible for firms to use smaller units of capital to produce their goods. This means that small firms cannot compete in the market.

A good example is the jumbo jet industry. The technological change which enabled very large jet airliners to be produced led to an outcome in which, in the western world, the American Boeing Corporation and the European Airbus consortium are the only two jumbo jet manufacturers. In Chapter 5, we will call this situation a **duopoly**.

How the process of creative destruction is linked to technological change

KEY TERM

creative destruction capitalism evolving and renewing itself over time through new technologies and innovations replacing older technologies and innovations.

The term **creative destruction** was first coined in 1942 by the Austrian economist Joseph Schumpeter to describe how capitalism, which dominates the economic system in which we live, evolves and renews itself over time. (Capitalism is the name given to the parts of the economy in which the means of production or capital are privately owned. In the UK, public limited companies or PLCs are the dominant form of capitalist business enterprise.)

In his book *Capitalism, Socialism, and Democracy*, Schumpeter wrote: 'The opening up of new markets, foreign or domestic … incessantly revolutionises the economic structure from within, incessantly destroying the old one, incessantly creating a new one. This process of Creative Destruction is the essential fact about capitalism.' Schumpeter also stated that 'The essential point to grasp is that in dealing with capitalism we are dealing with an evolutionary process.'

Creative destruction is strongly related to the processes through which technological change and innovation affect the ways in which businesses behave. It describes a process in which economic growth occurs in the economy as a result of new innovations creating more economic value than that being destroyed by the replacement of older technologies. Over time, societies that allow creative destruction to operate grow more productive and richer; their citizens benefit from new and better products and higher living standards. Creative destruction is central to the ways in which free-market economies and mixed economies develop and change over time.

APPLICATION OF ECONOMICS IN THE REAL WORLD 4.8

Apple and creative destruction

On 1 April 1976, Apple Computer Inc. was incorporated by three people: Steve Jobs, Steve Wozniak and Ron Wayne. Twenty-one years later in 1997, Steve Jobs, having left Apple following disputes about business strategy, rejoined the company and remained in charge until his death in 2011. (In 2007, Jobs had renamed the company Apple Inc. to reflect the fact that Apple had diversified away from computers into the iPod, the iPhone, the iPad and iTunes.)

Over this period, and particularly after 2001 when the iPod was first marketed, Apple had a crushing effect on specific competitors. According to Barry Ritholtz, writing in the *Washington Post* shortly after Jobs' death, this was creative destruction on a huge scale. Ritholtz argued that Jobs remade entire industries according to his unique vision. From music to film, mobile phones to media publishing and computing, Jobs' impact has been enormous.

Today, the triple threat of iPod/iPhone/iPad has left behind a wake of overwhelmed business models, confounded managements and bereft shareholders. The businesses which have been destroyed, or left as hollow versions of their former selves, include Hewett-Packard (HP), Nokia and Blackberry. According to Ritholtz, HP's printer business might still have some ink left in its cartridges, but its PC operations are hurting, gutted by sales of the iPad. HP's tablet entry, the TouchPad, was an unmitigated disaster, unable to compete with the iPad.

In 2007, the Finnish company Nokia totally dominated the mobile phone market. Many people thought that Nokia's lead was more or less insurmountable. But what has happened since is a reminder of just how quickly and completely the market power of a previously 'dominant' tech firm can disappear. Following the introduction of Apple's iPhone in 2007, both Nokia and Blackberry began a rapid decline.

Under a headline 'Once-cool Blackberry fails to keep pace with rivals', China's *Morning Post* described how Blackberry, an early mover in the high-end mobile phone market, lost market share mainly to Apple's iPhone and to smartphones powered by Google's Android operating system. While Blackberry was considered perhaps the trendiest if not the largest mobile phone maker several years ago, the company quickly lost momentum as it failed to keep pace with innovations from rivals. Gerry Purdy, an analyst at Compass Intelligence, said that 'The one gigantic issue facing Blackberry was the delay in getting into the smartphone market. And that was three years after the iPhone was released. So that's six years. The market was moving too fast.' Blackberry was too complacent, having become 'blinded' to competitive threats.

Even software giant Microsoft has suffered from Apple's innovation and marketing. Once Apple's main competitor in computer manufacturing and software, Microsoft has become vulnerable on multiple fronts. It has missed nearly every major trend in technology in recent years. Microsoft still has its cash cows Windows and its Office suite of products, but the company could lose out significantly to Apple in the next few years.

Follow-up questions

1. What is meant by 'creative destruction'?
2. This case study was written at the time of Steve Jobs' death in 2011. Find out what has happened to Nokia's and Blackberry's smartphone businesses in the years since then.
3. Cathode-ray television-tube manufacturers, video rental shops, high-street travel agents and bookshops have all in recent years been victims of creative destruction. Research how this has happened in one of these industries and explain why you think consumers may or may not have benefited from the process.
4. The marketing of smartphones by ICT companies such as Apple has contributed to the growth of social media. Explain how social media have been affecting more traditional forms of media such as newspapers and television.

SECTION 4.8 SUMMARY

- The ways in which firms operate are affected by technological change, which encompasses the processes of invention, innovation and diffusion of technology in the economy.
- Technological change involves changing existing technologies, hopefully for the better, and developing completely new technologies, both to improve existing products and the processes involved in making the products, and to develop completely new products and processes.
- Technical progress means applying scientific and engineering knowledge, as it develops, to produce goods which are more efficient and work better.
- Innovation improves on or makes a significant contribution to something that has already been invented, thereby turning the results of invention into a product.
- Technological change affects methods of production, productivity, efficiency and firms' costs of production.
- Technological change has been very significant in the development of new products and new markets, and the destruction of existing markets.
- Creative destruction is a process through which capitalism evolves and renews itself over time through new technologies and innovations replacing older technologies and innovations.

Questions

1 Why is it important first to understand production theory in order to understand cost theory?

2 Distinguish between the short run and the long run.

3 Explain the mathematical relationship between the marginal and the average values of a variable.

4 State the law of diminishing returns.

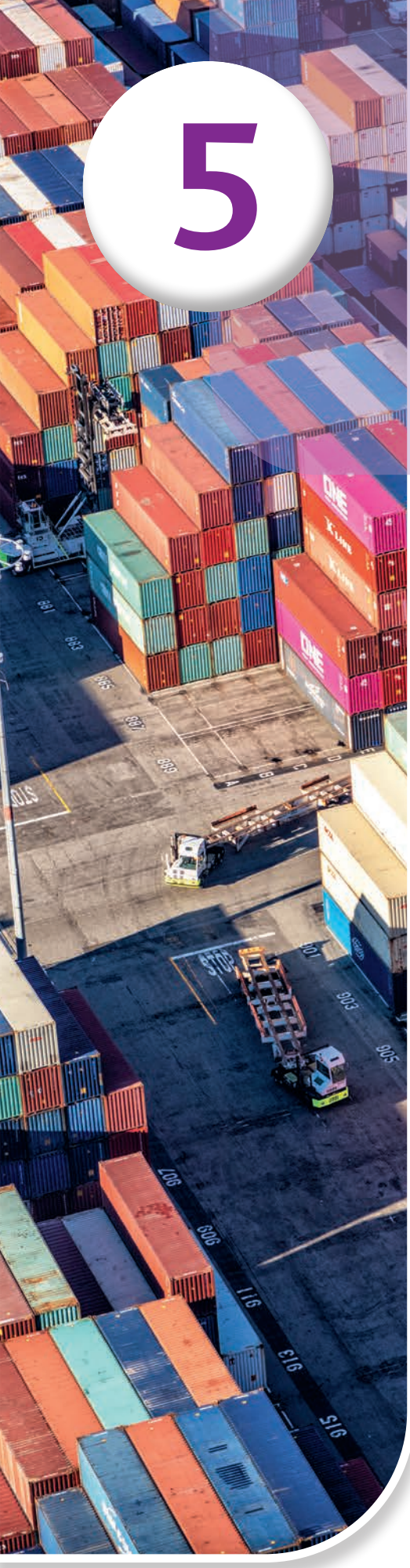
5 What is the difference between the marginal returns to a variable factor of production and returns to scale?

6 Distinguish between internal and external economies of scale.

7 Why are a firm's revenue curves dependent on the competitiveness of the market structure in which it exists?

8 Define the term 'profit'.

9 Distinguish between technological change and technical progress.



5 Perfect competition, imperfectly competitive markets and monopoly

Chapter 4 has already introduced you to the two market structures of perfect competition and monopoly, and briefly mentioned the imperfectly competitive market structures of monopolistic competition and oligopoly that lie between the two extremes on the market structure spectrum. This chapter draws on the information about cost and revenue curves provided in Chapter 4 to explain profit maximisation in perfect competition and monopoly. After evaluating these two market structures using efficiency criteria, the chapter then looks at the two intermediate market structures of monopolistic competition and oligopoly. The order in which we cover the different market structures differs from how they are set out in the A-level specification. This is because we believe that it is easier to appreciate the nature of monopolistic competition and oligopoly after first understanding monopoly. The chapter concludes with an explanation of the pricing policies, such as price discrimination, that firms may use in imperfectly competitive market structures, followed by comparison of perfect competition and monopoly markets in terms of the welfare criteria of consumer surplus and producer surplus.

LEARNING OBJECTIVES

These are to understand:

- profit maximisation in different market structures
- whether perfect competition is more efficient than monopoly
- how monopolistic competition and oligopoly are forms of imperfect competition
- the use of concentration ratios to define oligopoly in terms of market structure
- oligopoly in terms of market behaviour or conduct, in the context of interdependence among oligopolists
- the difference between competitive and collusive oligopoly
- why and how firms undertake price discrimination
- the different forms of economic efficiency
- how dynamic efficiency relates to the process of creative destruction and how firms compete in real-world markets
- the welfare criteria of consumer surplus and producer surplus and how they can be used to evaluate the different market structures.

5.1 Market structures

KEY TERM
market structure the organisational and other characteristics of a market.

Market structure can be defined in terms of the organisation and other characteristics of a market. Important features of market structure include:

- the number of firms in the market
- the market share of the largest firms, which as we later explain can be measured with the use of concentration ratios
- the nature of the costs incurred by the firms in the market
- the nature of the sales revenue earned by firms in the market
- the extent to which there are barriers to entry to, and exit from, the market
- ease of access to information about what is going on in the market
- the extent to which firms in the market undertake product differentiation and adopt different price-setting procedures
- the ways, if any, in which firms are affected by buyers' behaviour in the market.

The spectrum of competition

In the context of markets, the word 'spectrum' encompasses the range of market structures which lie between the two extremes of perfect competition and monopoly. Figure 5.1 is similar to the lower part of Figure 4.1 in Chapter 4. However, Figure 5.1 contains more information about each of the market structures that we explain in this chapter.

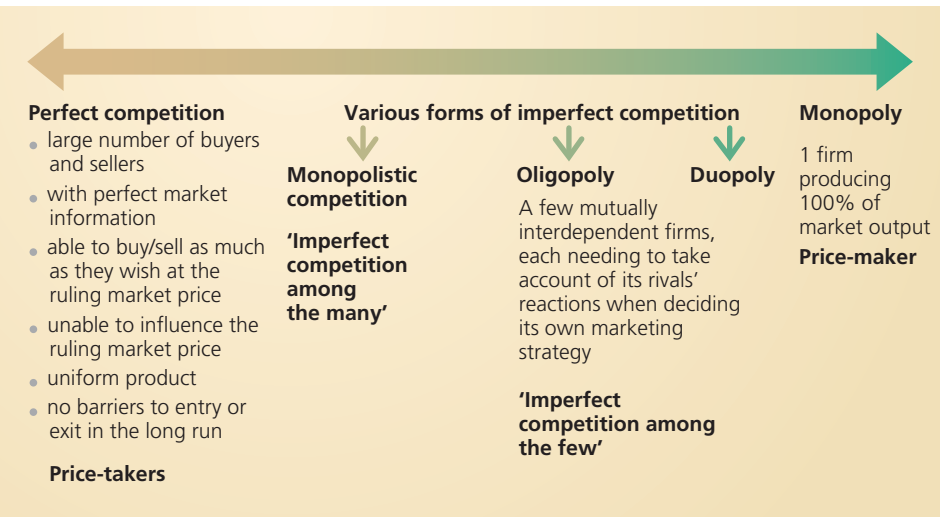


Figure 5.1 The spectrum of market structures

Distinguishing between different market structures

The number of firms in a market

Factors such as the number of firms in the market, the degree of product differentiation and ease of entry are used to distinguish between different market structures. A requirement for a market to be perfectly competitive is a large number of buyers and sellers, while at the other extreme pure monopoly is defined by the fact that there is just one firm in the market.

TEST YOURSELF 5.1
Explain the differences between perfect competition and monopoly.

Perfect competition is in fact an abstract theoretical extreme rather than a real-world market structure. It is impossible for markets to display simultaneously all the conditions, listed in Figure 5.1, which are necessary for perfect competition to exist. Since any violation of the conditions of perfect competition immediately renders a market imperfectly competitive, even the most competitive markets in the real economy are examples of imperfect competition rather than perfect competition. The best we can say is that some of the highly competitive markets in the world we live in approximate to perfect competition, but nevertheless they are not perfectly competitive.

At the other extreme, pure monopoly does exist, though it is more accurate to say that market structures *dominated* by one firm are much more common than pure monopoly. In the UK, water companies provide examples of pure monopoly. Depending on where you live, there is only one company supplying tap water to your house or flat. For example, if you live in central London, you must buy your tap water from Thames Water — it is impossible to shop around and buy your water from an alternative supplier. You could, of course, buy bottled water instead of tap water, but this would not be a realistic choice for non-drinking uses of water, such as taking a shower or washing the dishes. And even if in other industries a firm appears to be a pure monopoly, the availability of substitute products and foreign competition through which overseas firms compete with the dominant domestic firm weaken the firm's monopoly power.

Almost all real-world firms are therefore better described as imperfectly competitive, located between the two extremes of perfect competition and pure monopoly. As Figure 5.1 indicates, monopolistic competition is often described as imperfect competition among the many, which means that, as in perfect competition, there are a large number of firms in such markets. However, unlike perfect competition, many real-world markets display the characteristics of monopolistic competition which we explain later in this chapter. Typical examples are high-street coffee shops and newsagents.

We will also examine oligopoly later in the chapter. In oligopoly, there are only a few firms in the market, at least in terms of large firms, though often the large firms coexist with a number of smaller firms. Figure 5.1 describes oligopoly as 'imperfect competition among the few'. As we explain later in the chapter, concentration ratios are used when oligopoly is defined in terms of market structure, while interdependence among the firms that make up the market is a main characteristic when oligopoly is defined in terms of market behaviour.

Market entry and exit barriers

KEY TERM
entry barriers obstacles that make it difficult for a new firm to enter a market.

Market structure is also affected by the ease of entry into (and exit from) the market. In the short run, when at least one factor of production (usually assumed to be capital) is fixed, firms cannot enter or leave the market, whatever the market structure. In the long run, by contrast, when all the factors of production are variable, firms can enter or leave competitive markets. However, at the other end of the market spectrum, pure monopoly is protected by **entry barriers** in the long run as well as in the short run, while in an oligopolistic market there may still be significant entry barriers in the long run.

KEY TERMS

incumbent firm a business with an established position in the market which often benefits from consumer loyalty and economies of scale.

natural barriers (also known as **innocent barriers**) barriers to market entry which are not caused by deliberate actions of firms in the industry.

sunk costs costs that have already been incurred and cannot be recovered.

artificial barriers 'man-made' barriers to market entry (e.g. patent protection).

An entry barrier is a cost of production which must be borne by a firm that seeks to enter an industry, but which is not borne by **incumbent firms**, which are already in the market. In pure monopoly, entry barriers are permanent in the sense that if they are removed, the entry of new firms, attracted by the monopolist's abnormal profit, immediately means the monopoly is destroyed. (At this point, it is worth noting that an efficient monopolist, making normal profit, will not attract new entrants into the market.) However, building on the concepts of technological change and creative destruction introduced in Chapter 4, later in this chapter we will be explaining how the development of competing new products and technologies weakens and often destroys the monopoly power of dominant firms whose power was previously seen as impossible to overcome.

There are two main types of entry barrier: natural barriers and artificial (or man-made) barriers.

Natural barriers

Natural barriers to market entry, which are also known as **innocent barriers**, include:

- **Economies of scale.** Established large firms produce at a lower long-run average cost, and are more productively efficient, than smaller new entrants, who become stranded on high-cost short-run average cost curves.
- **Indivisibilities.** These are an example of technical economies of scale, preventing certain goods and services being produced in plants below a certain size. Indivisibilities occur, for example, in metal smelting and oil refining industries.
- **Sunk costs.** These are costs which cannot be recovered if a firm decides to leave a market, and which therefore increase the risk of entry. See page 158 for more on sunk costs.

Artificial barriers

In contrast, **artificial barriers**, which are also known as *strategic barriers*, are the result of deliberate action by incumbent firms to prevent new firms from entering the market. Strategic entry barriers include:

- **Patents.** These provide legal protection for an invention and for all the variants of a product that develop from it. A patent grants the holder the sole right, effectively a legal monopoly, to make, use or sell a specific product. In the UK, this right can last for up to 20 years.
- **Product differentiation.** By differentiating their products, which then become protected by intellectual property and trade mark legislation, firms protect themselves from 'copy-cat' market entrants.
- **High levels of expenditure on advertising and marketing.** Established firms can make it difficult for new competitors by spending heavily on advertising and marketing, which are irrecoverable expenditures and a form of sunk cost if the firm decides to leave the market.
- **Benefiting from 'first mover' advantage.** By being first into a market, 'first movers' such as Apple can establish themselves, build a customer base and make it difficult for later arrivals to compete.
- **Limit pricing and predatory pricing.** Limit pricing occurs when firms already in the market reduce prices so that they only just make normal profit, in order to deter or limit the entry of new competitors. Predatory pricing goes one stage further — occurring when an established or incumbent firm deliberately sets prices below costs to force new market entrants out of business.

KEY TERM
exit barriers obstacles that make it difficult for an established firm to leave a market.

Closely related to entry barriers are **exit barriers**, which make it difficult for an incumbent firm to leave a market. Exit barriers may include costs associated with redundancy payments to workers and the reduction in value of stock which might have to be sold off quickly. Sunk costs may also act as an exit barrier.

TEST YOURSELF 5.2

Explain two examples of (a) natural barriers to market entry; (b) artificial barriers to market entry; and (c) sunk costs.

EXTENSION MATERIAL

Limit pricing and predatory pricing

Firms may set the prices at which they sell their goods to create an artificial barrier to market entry. This is when they set limit prices or, in a more extreme case, predatory prices. When natural barriers to market entry are low or non-existent, incumbent firms (i.e. firms already in the market) may set low prices, known as limit prices, to deter new firms from entering the market. Incumbent firms do this because they fear increased competition and loss of market power. With limit pricing, firms already in the market sacrifice short-run profit maximisation in order to maximise long-run profits, achieved through deterring the entry of new firms.

Should limit pricing be regarded as an example of a competitive pricing strategy, which reduces prices and the supernormal profits enjoyed by the established firms in the market? Or is limit pricing basically anti-competitive and best regarded as an unjustifiable restrictive practice? The answer probably depends on circumstances, but when limit pricing extends into predatory pricing, there is a much clearer case that such a pricing strategy is anti-competitive and against consumers' interest.

Whereas limit pricing deters market entry, successful use of predatory prices removes recent entrants to the market. Predatory pricing occurs when an established or incumbent firm deliberately sets prices below costs to force new market entrants out of business. Once the new entrants have left the market, the established firm may decide to restore prices to their previous levels.

Product differentiation

In Chapter 4, we defined a firm as a productive organisation that sells its output of goods or services commercially. Many decades ago, when economists were first putting together the theory of the firm and developing the model of perfect competition, not only did they regard firms as one-product business organisations, they assumed that within a market, all firms produce a uniform or homogeneous product. (This is one of the conditions of perfect competition listed in Figure 5.1.)

KEY TERM
product differentiation the marketing of generally similar products with minor variations or the marketing of a range of different products.

While there are a large number of small firms in the UK economy today, relatively few produce just a single good or service. Most firms — particularly large and medium-sized business enterprises, but also small businesses — undertake varying degrees of **product differentiation**. Firms often produce a range of relatively similar products, some of which compete with each other, but others of which are aimed at differentiated market segments. Mobile phones provide examples of both. Apple is well known for introducing two new smartphones roughly every year: a 'high-end' and a slightly more basic (and cheaper) model. However, Apple continues to manufacture and sell 'last year's model'. As a result, the latest models — at the time of writing, the iPhone 14 models — compete with earlier models which Apple still

sells, such as the iPhone 12 and the iPhone 13. Samsung, which is Apple's main rival in the smartphone market, differentiates its phones in a similar way, but launches its new phones more frequently and in a wider variety of options than its competitor.



Apple iPhones are an example of successful product differentiation

SECTION 5.1 SUMMARY

- Market structures provide the framework in which businesses exist.
- Different market structures display different degrees of competitiveness.
- There are two main types of entry barrier making it difficult or preventing a new firm from entering a market: natural barriers and artificial (or man-made) barriers.
- Sunk costs cannot be recovered if a firm decides to leave a market.
- Firms of all sizes undertake varying degrees of product differentiation.

5.2 The objectives of firms

Profit maximisation

We have already mentioned how traditional theory assumes that the owners and entrepreneurs who run firms have only one business objective: to produce the level of output at which profit is maximised.

For all firms, whatever their business objective(s):

total profit = total revenue – total cost

Profit maximisation requires that a firm produces the level of output at which *TR* – *TC* is maximised, whatever the market structure in which the firm produces and sells its output. Given the profit-maximising objective, if the

firm succeeds in producing and selling the output yielding the biggest possible profit, it has no incentive to change its level of output.

The profit-maximising rule (*MR = MC*)

It is generally more useful to state the condition for profit maximisation as:

marginal revenue = marginal cost, or *MR* = *MC*

MR = *MC* means that a firm's profits are greatest when the addition to sales revenue received from the last unit sold (marginal revenue) equals exactly the addition to total cost incurred from the production of the last unit of output (marginal cost).

Imagine, for example, a market gardener producing tomatoes for sale in a local market, but unable to influence the ruling market price of 50p per kilo. At any size of sales, average revenue is 50p, which also equals marginal revenue. Suppose that when the horticulturalist markets 300 kilos of tomatoes, the cost of producing and marketing the 300th kilo is 48p. If the tomato grower decides not to market the 300th kilo, 2p of profit is sacrificed. Suppose also that total costs rise by 50p and 52p respectively when a 301st kilo and a 302nd kilo are marketed. The sale of the 302nd kilo causes profits to fall by 2p, but the 301st kilo of tomatoes leaves total profits unchanged: it represents the level of sales at which profits are exactly maximised.

To sum up:

- when *MR* > *MC*, profits rise when output increases
- when *MR* < *MC*, profits rise when output reduces.

So only when *MR* = *MC*, at the level of output Q_1 in Figure 5.2, are profits maximised.

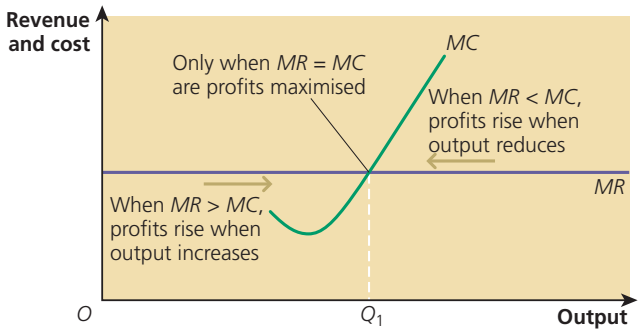


Figure 5.2 A firm's profit-maximising output occurs when *MR* = *MC*

When *MR* > *MC* or *MR* < *MC* the firm fails to maximise profit. To maximise profit, the firm must change its level of output until it reaches the point at which *MR* = *MC*. Once this is reached, the firm has no incentive to change output, unless some event disturbs either costs or revenues.

It is important to understand that firms in *all* market structures (perfect competition, monopoly and imperfectly competitive markets such as monopolistic competition and oligopoly) can only maximise profit when marginal revenue equals marginal cost. *MR* = *MC* is a universal condition that must be met for profit maximisation to occur, whatever the market structure.

TEST YOURSELF 5.3

Distinguish between profit maximisation and revenue maximisation.

The divorce of ownership from control

In a firm, the entrepreneur is the decision-maker and financial risk-taker, providing answers to such standard economic questions as *what*, *how*, *how much*, *where* and *when* to produce. In many small firms, the owner of the

KEY TERM

divorce of ownership from control the owners and those who control the firm (managers) are different groups with different objectives.

TEST YOURSELF 5.4

Who owns a UK public limited company (PLC)?

business is the entrepreneur, so ownership and control lie with the same person. But this is generally not true for larger businesses, where there is a **divorce of ownership from control**. Medium-sized and large companies are often owned by thousands of shareholders, though the majority of shares are usually owned by a relatively small number of financial institutions such as pension funds and insurance companies. However, management decisions are made by executive directors, who are members of the company's corporate board and employed by the company, and by other managers or executives whom they in turn employ.

In theory, the directors of a public company who exercise the entrepreneurial function are answerable to the shareholders. This means that, in the event of bad performance and a failure to maximise profit, the directors can be voted out of office. In practice, however, this seldom happens.

How divorce of ownership from control may affect firms' objectives, conduct and performance

Perhaps the most important problem resulting from the divorce between ownership and control is the possibility that directors and managers will pursue an agenda of their own, which is not in the interests of the shareholders as a body. This is an example of what is known as the principal–agent problem, or agency problem, where shareholders are the principals and managers are the agents. The problem stems from the fact that shareholders, who in a company are the owners of the business, may have an objective (profit maximisation) which differs from the objectives of the managers they employ. Rather than maximise profits, managers or business executives, who possess a monopoly of technical knowledge about the actual running of the company, concentrate on achieving managerial objectives such as maximising their career prospects or creature-comforts.

Conflicts of interest arise between agents and principals when the incentives which affect their behaviour are not the same. The agent bears the cost of fulfilling the task delegated by the principal, but the agent does not usually receive the full benefit of their actions. This destroys the incentive for the agent to put the same effort into the task that would be the case if the agent were acting solely on their own behalf. Related to this is the fact that agents may take on excessive risk if they enjoy the benefits of doing so, but not the costs. For example, when managers take on excessive risk, in the event of success they are rewarded with high bonuses, but in the event of failure the owners of the business end up losing a lot of money. This illustrates the problem of 'moral hazard', which occurs when one person takes more risks because someone else has agreed to bear the burden of those risks.

There are two important reasons why an agent can get away with not acting in the best interest of the principal. First, the cost to the principal of sacking or punishing the agent may be too high relative to any benefit the principal will enjoy. Secondly, information asymmetry may result from the fact that the agent knows more than the principal about what is going on in the business. It is often difficult or even impossible for shareholders to know whether the directors and managers they employ are acting in their best interests. For example, if a company reports disappointing profits, shareholders may find it difficult to judge whether the blame stems from managerial laziness and incompetence, or whether the poor results are due to adverse economic factors outside management control.

Various methods can be used to realign the incentives facing the owners of a business and the directors and managers they employ. These include profit-related pay and paying managers partly by giving them company shares (executive stock options). These are financial incentives, but threats such as dismissal can also be used. (However, the threat of dismissal is somewhat weakened by the fact that all too often unsuccessful managers are 'rewarded for failure' with large financial pay-offs when they leave the business.)

Possible business objectives other than profit maximisation

In the context of the principal–agent problem, we have already mentioned the pursuit by the managers who run a business of objectives such as maximising management status and power. Another business objective is to provide good customer service and quality of product, both in terms of the product itself and in the provision of good-quality and reliable after-sales service. These objectives, which are particularly important for socially minded owners and managers, are not in themselves inconsistent with the profit-maximising objective. Very often they are best seen as means towards an end, the end being the achievement of some other business objective such as profit maximisation.

Growth maximisation and survival

'Growth for growth's sake' is another possible business objective, though usually the continuing growth of a business is best viewed as a means of achieving economies of scale and their associated cost reductions. Growth can contribute to managerial prestige. It can also be a means of achieving monopoly power by knocking out smaller competitors. In the case of the proprietors of newspapers and television stations, growth is a means of increasing political influence.

At the other extreme, loss-making businesses, and those that fear their low level of profits might turn into losses, often see survival as a primary business objective. This is most common when the economy is hit by recession. Businesses hope they can ride out the adverse effects of economic downturn, survive and continue to grow once the worst is over.

Sales revenue maximisation

Firms may also try to maximise sales revenue, particularly when managerial pay is linked to revenue rather than profit. Students often confuse profit maximisation with revenue maximisation, but the two concepts are different. As we have explained, profit maximisation occurs at the level of output at which the difference between a firm's total sales revenue (*TR*) and its total costs of production (*TC*) is greatest. This is also the level of output at which marginal revenue equals marginal cost (*MR* = *MC*). By contrast, revenue maximisation occurs at the level of output at which marginal revenue is zero.

The difference between profit maximisation and revenue maximisation is shown in Figure 5.3. The profit-maximising level of output Q_1 is located below point *X* where *MR* = *MC*. By contrast, the revenue-maximising level of output Q_2 is located at point *Z*, where *MR* = 0. Providing the *AR* and *MR* curves slope downwards to the right, the profit-maximising level of output is always below the revenue-maximising level of output.

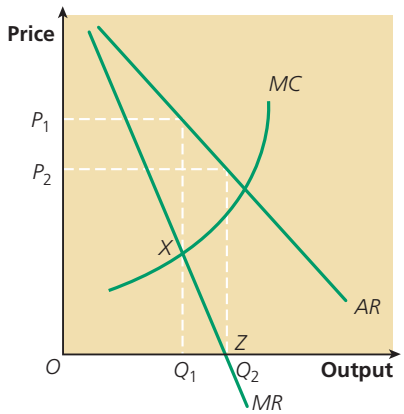


Figure 5.3 Profit maximisation and revenue maximisation when a firm's *AR* curve slopes downwards

STUDY TIP

Make sure you don't confuse *profit maximisation* with *revenue maximisation*. Note also that output Q_2 at point Z in Figure 5.3 is located immediately below the point on the *AR* curve where price elasticity of demand is neither elastic nor inelastic (i.e. it is unit elastic). In Figure 5.3, the profit-maximising level of output, Q_1, lies below the elastic section (top half) of the *AR* curve.

QUANTITATIVE SKILLS 5.1

Worked example: profit maximisation

At the firm's present level of output, an entrepreneur determines that its marginal cost is £36 (and on the rising portion of the curve), and its marginal revenue is £42 (on a downward-sloping *MR* curve).

The firm will maximise profits or minimise losses by:

A increasing price while keeping output constant
B decreasing price and increasing output
C decreasing both price and output
D increasing both price and output

The correct answer is B. Since *MR* > *MC* at the firm's *current* level of output, profits rise when the firm increases its level of output. The question tells us that the firm has a downward-sloping *MR* curve. This means that its *AR* curve also slopes downwards. Since *AR* is the same as price, whenever the firm increases output, it has to accept a lower price. (If you refer back to Figure 5.3, the firm would initially be producing an output level below and to the left of Q_1. By increasing output to Q_1, price would fall and profit would increase up to Q_1, but fall thereafter.)

The satisficing principle

As we saw in Chapter 2, behavioural economists argue that it is too crude to view individuals and firms simply as utility and profit maximisers. With regard to the nature of production, behavioural economists see the firm as an organisation comprising coalitions of different groups within the firm, each possessing different group objectives. The different groups, often called *stakeholders* in the business, may each have a different vested interest in terms of how the business performs.

KEY TERM

satisficing achieving a satisfactory outcome rather than the best possible outcome.

Different stakeholder groups have different views on what the company should be doing. Managers, for example, may seek prestige, power and high salaries. Besides wanting higher wages, workers may be looking for improved job security and working conditions. In their desire to achieve higher profits, shareholders may try to keep down wages and managerial salaries. Differing goals or aspirations can result in group conflict. Management may try to resolve conflict between the different interest groups by replacing the profit-maximising objective with **satisficing**. Satisficing, which means achieving a satisfactory outcome rather than the best possible outcome, may help to resolve the conflict between managers' and shareholders' objectives. While trying to maximise executive creature-comforts such as managerial status, salaries, fringe benefits and career structures, a company's board of directors must keep shareholders happy. According to this theory, managers maximise their own objectives, but subject to the constraint of delivering a satisfactory level of profit for shareholders.

However, attempting to satisfice so as to satisfy the aspirations of as many groups within the firm as possible means compromise and the possible setting of minimum rather than maximum targets. Satisficing is particularly likely for monopolies and firms in imperfectly competitive markets protected by entry barriers. In these circumstances, in seeking an easy life, a firm's managers may content themselves with satisfactory profit, combined with a degree of inefficiency or unnecessary costs, which could in principle be eliminated. In highly competitive markets, by contrast, firms that are initially content with satisfactory profits and a degree of inefficiency might be forced by the entry of

TEST YOURSELF 5.5

Explain why one of the five possible answers to the following question is correct and why the other four answers are wrong.

A firm engaged in satisficing behaviour is most likely to:

A maximise profits
B maximise revenue
C produce at an output different from that of a profit-maximising firm
D maximise sales
E minimise costs

new, more competitive firms to eliminate unnecessary costs. In order to survive, firms end up 'as if' they are profit maximisers, even though they may appear, at first sight, to be satisficing or pursuing some other business objective.

SECTION 5.2 SUMMARY

- Traditional theory assumes that the owners and entrepreneurs who run firms have only one business objective: to produce the level of output at which profit is maximised.
- Profit is maximised at the level of output at which *MR* = *MC*.
- In large businesses, there is often divorce of ownership from control.
- Firms may have business objectives other than profit maximisation, such as growth maximisation and sales maximisation.
- Firms may be satisficers rather than maximisers, aiming for satisfactory rather than maximum profits.

5.3 Perfect competition

Mention has been made of perfect competition on several occasions both in this chapter and in Chapter 4. To recap, perfect competition does not actually exist in real-world economies because it is impossible to meet simultaneously all the six conditions listed in Figure 5.1. Nevertheless, perfect competition plays a significant role in the traditional theory of the firm because it provides a yardstick against which the desirable and undesirable properties of real-world markets can be measured. For example, we will use perfect competition as a benchmark against which to judge whether monopoly functions efficiently or inefficiently, and the extent to which resource misallocation occurs in monopoly.

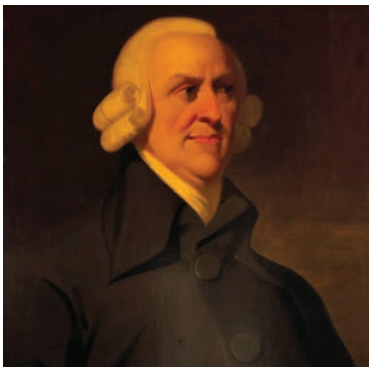
APPLICATION OF ECONOMICS IN THE REAL WORLD 5.1

The 'invisible hand' of the market

In 1776, in his book *An Inquiry into the Nature and Causes of the Wealth of Nations*, the economist Adam Smith wrote:

> An individual neither intends to promote the public interest, nor knows how much he is promoting it. He intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love.

Smith was making the argument, repeated ever since by economists of a free-market persuasion, that the pursuit of individual self-interest in the market economy leads to outcomes which are in the common good or public interest — as long as the markets are free and competitive. Consumers benefit from individualistic behaviour in competitive markets because prices and profits end up being lower than would be the case if the markets were dominated by a few large firms.



Adam Smith argued consumers would benefit from producers in competitive markets following their own interests

Follow-up questions

1. Outline one way in which consumers can benefit from a lack of competition in a market.
2. Describe one circumstance, apart from when there is little or no competition in a market, in which the pursuit of self-interest may be undesirable in the economy.
3. When may cooperation between individuals be preferable to individualistic economic behaviour?
4. Does the 'invisible hand' of the market always lead to markets performing well?

EXTENSION MATERIAL

The importance of self-interest

Economists generally regard competitive markets as desirable. However, the desirable properties of competitive markets (namely economic efficiency, welfare maximisation and consumer sovereignty) do not result from any assumption that business people or entrepreneurs in competitive industries are more highly motivated or public spirited than monopolists. Traditional economic theory assumes that everyone is motivated by self-interest and by self-interest alone. This applies just as much to firms in competitive markets as it does to monopolies. Entrepreneurs in competitive industries would very much like to become monopolists, both to gain an easier life and also to make bigger profits. Indeed, from a firm's point of view, competing successfully means eliminating competition and becoming a monopoly. But in perfect markets, market forces (Adam Smith's invisible hand of the market) and the absence of barriers to entry and exit prevent this happening.

Imagine, for example, a situation in which a firm in a perfectly competitive industry makes a technical breakthrough which reduces production costs. For a short time the firm can make significant profits. But because, in highly competitive markets, market information is available to all firms, other firms within the market and new entrants attracted to the market can also enjoy the lower production costs.

Ultimately, of course, consumers benefit from lower prices brought about by technical progress and the forces of competition, but it is market forces, and not some socially benign motive or public spirit on the part of entrepreneurs, that accounts for the optimality of a highly competitive market as a market structure.

APPLICATION OF ECONOMICS IN THE REAL WORLD 5.2

How the use of mobile phones can make markets more competitive

For a market to be competitive, buyers and sellers need accurate information about supply and demand. Before the use of mobile phones, fishermen in southern India lacked information about prices being charged for newly caught fish in other fishing villages along the coast. This lack of adequate information about conditions of supply and prices being charged led to small, separated and relatively uncompetitive fish markets.

If a fisherman made a good catch, other fishermen operating out of his home port and fishing in the same area would also catch a lot of fish. But when all the fishing vessels sailed back home, fish prices in the local village fish market would slump because of excess supply.

Another possibility was to sail down the coast after the catch was made, in the hope that in other villages fish catches were less bountiful and prices were therefore better. But, because of high fuel prices and uncertainty about what might be happening elsewhere, fishermen generally chose to return to their own village. This was wasteful because over-supply led to fish being thrown away, even though they might have been sold in slightly more distant fish markets. Another result was that there were wide variations in fish prices in different fishing villages.

However, after mobile phones had been introduced in southern India, while they were still at sea fishermen began to call markets all along the coast to find out where prices were highest. Having obtained this information, fishermen were now prepared to market their fish further afield, despite the fuel costs involved. The number of unsold fish that previously had been thrown back into the sea fell dramatically. Fish prices also fell. The 'law of one price' was operating — there was now a single price along the coast for more or less identical fish. By improving the exchange of information between fishermen, mobile phone technology has therefore contributed to the growth of a larger and much more competitive market.

Follow-up questions

1 Explain how this account illustrates how better information on the part of buyers or sellers improves the way a market functions.
2 Name two UK markets that have been made more competitive as a result of the growing use of mobile phones.
3 Describe one economic disadvantage resulting from the use of mobile phones.
4 What is the 'law of one price' mentioned in the passage?

Short-run profit maximisation in perfect competition

At this stage, you should refer back to Chapter 4, page 105, and look again at Figure 4.19. The two graphs in the diagram illustrate how each firm in a perfectly competitive market passively accepts the ruling market price, which becomes each firm's average revenue (*AR*) and marginal revenue (*MR*) curve. The third condition of perfect competition tells us that a perfectly competitive firm can sell as much as it wishes at the market's ruling price. But how much will it actually wish to produce and sell? If we assume that each firm's business objective is solely to maximise profit, the answer is shown in Figure 5.4.

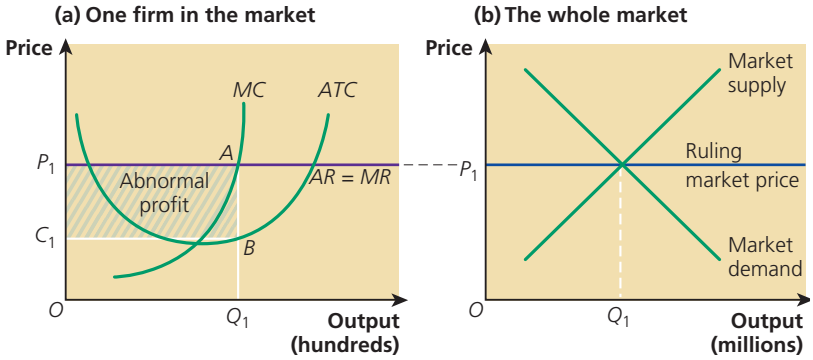


Figure 5.4 Short-run profit maximisation in perfect competition

As in Figure 4.19, Figure 5.4(b) shows how the ruling market price (P_1) is determined at the intersection of the demand and supply curves in the market as a whole. When beamed horizontally into Figure 5.4(a), the ruling market price becomes the average revenue (*AR*) curve and the marginal revenue (*MR*) curve facing each firm in the market. However, in Figure 5.4(a), we have added the perfectly competitive firm's average total cost (*ATC*) and its marginal cost (*MC*). Point *A* in Figure 5.4(a) (at which *MR* = *MC*) locates the profit-maximising level of output Q_1, with the firm's output shown in hundreds of units. At this level of output, total sales revenue is shown by the area OQ_1AP_1 (output multiplied by average revenue or price). Total cost is shown by the rectangular area OQ_1BC_1 (output multiplied by average cost). Abnormal profits (measured by subtracting the total cost rectangle from the total revenue rectangle) are shown by the shaded area C_1BAP_1.

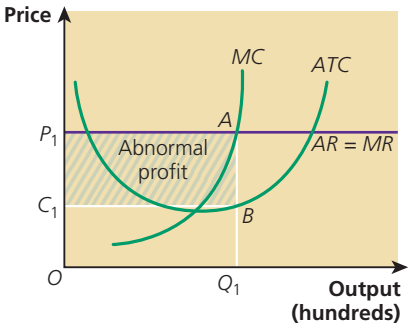


Figure 5.5 A perfectly competitive firm maximising profit in the short run

In Figure 5.5, the graph is the same as Figure 5.4(a), but presented without any information about what is going on in the market as a whole. The diagram enables you to focus on a firm's short-run profits in conditions of perfect competition, especially the positions of the cost and revenue curves and the abnormal profit rectangle.

TEST YOURSELF 5.6

Explain the difference between the short run and the long run.

Long-run profit maximisation in perfect competition

Referring back to the list of the conditions of perfect competition, you will see that although firms cannot enter or leave the market in the short run, they can do so in the long run. Suppose that in the short run, firms make abnormal profit, as just illustrated in Figures 5.4 and 5.5. In this situation, the ruling market price signals to firms outside the market that abnormal profits can be made, which provides an incentive for new firms to enter the market.

Figure 5.6 shows what might happen next. Initially, too many new firms enter the market, causing the supply curve to shift to the right to S_2 in Figure 5.6(b). This causes the price line to fall to P_2, which lies below each firm's *ATC* curve. When this happens, firms make a loss (or *subnormal profit*). However, just as abnormal profit creates the incentive for new firms to enter the market, subnormal profit has the opposite effect of causing some firms to leave the market.

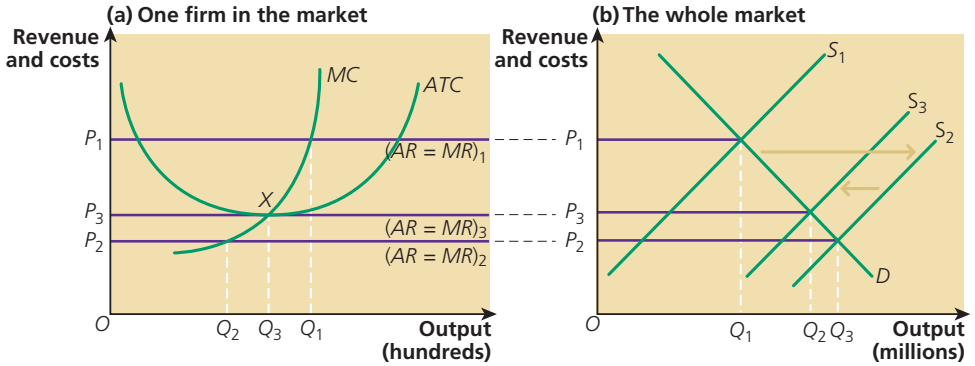


Figure 5.6 Perfect competition in the long run

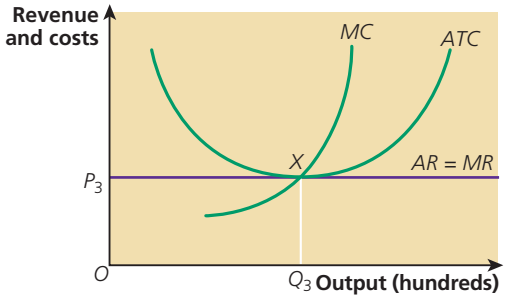


Figure 5.7 A perfectly competitive firm making only normal profit in the long run

When the price line lies below the *ATC* curve in Figure 5.6(a), and firms start leaving the market, the market supply curve shifts to the left in Figure 5.6(b). This causes the market price to rise. Eventually, the price settles at P_3, where, as Figure 5.6(a) shows, surviving firms make normal profit only. In this situation, each firm produces the level of output, measured in hundreds of units, immediately below point *X* in the diagram.

Figure 5.7 is similar to Figure 5.6(a), but only shows the price line P_3. In the diagram, a firm's total revenue and also its total cost are shown by the rectangle OQ_3XP_3. The entry of new firms into the market, attracted by short-run abnormal profits, has whittled away these profits to produce a long-run outcome in which surviving firms make normal profit only. (Remember, normal profit is treated as a cost of production, and is not shown explicitly in the diagram.) Since only normal profit is made, there is no incentive for firms to enter or leave the market.

TEST YOURSELF 5.7

A perfectly competitive firm can be described as a 'passive price-taker'. Explain what this means. What would happen in the short run and in the long run if a perfectly competitive firm tried to sell its output at a price above the ruling market price for the good the firm is producing?

Perfect competition and the efficient allocation of resources

On page 114 we stated that there are a number of types of economic efficiency. One of those which we identified is **productive efficiency**, which we first introduced in Chapter 1. We explained that for a particular firm in the market, the productively efficient level of output is the one immediately below the lowest point on the firm's *ATC* curve, at which the average cost per unit of output is minimised.

KEY TERM

productive efficiency for the economy as a whole occurs when it is impossible to produce more of one good without producing less of another. For a firm it occurs when the average total cost of production is minimised.

KEY TERM

allocative efficiency occurs when it is impossible to improve overall economic welfare by reallocating resources between markets. In the whole economy, price must equal marginal cost ($P = MC$) in every market.

In this chapter we introduce another type of efficiency, **allocative efficiency**, and discuss the extent to which perfect competition is both productively and allocatively efficient. Later in the chapter, we will apply the same analysis to the three other market structures you need to know: monopoly, monopolistic competition and oligopoly.

Productive efficiency and perfect competition

Figure 5.7 on page 132 shows the productively efficient level of output in perfect competition. This is output Q_3, which is located immediately below point *X* at the lowest point on the *ATC* curve. This is the average cost minimising level of output.

SYNOPTIC LINK

We introduced to you the concept of productive efficiency in Chapter 1. Refer back to page 133, which explains productive efficiency for the whole economy. When the economy is producing on its production possibility frontier, it is impossible to produce more of one good without producing less of another. All points on the economy's production possibility frontier are productively efficient.

Allocative efficiency and perfect competition

Allocative efficiency occurs when it is impossible to improve overall economic welfare by reallocating resources between industries or markets (assuming an initial distribution of income and wealth). As we will shortly explain under the heading 'A further look at allocative efficiency' (page 134), allocative efficiency is achieved when price equals marginal cost ($P = MC$) in every market in the economy.

Now look again at Figure 5.7. This shows that when a perfectly competitive market is in long-run equilibrium, the price the consumer pays (P_3) does indeed equal the marginal cost of production. In long-run equilibrium, a perfectly competitive firm is both productively and allocatively efficient, though when economies of scale are possible, which perfectly competitive firms cannot achieve, this is no longer the case.

But will perfect competition achieve an efficient allocation of resources?

To recap, Figure 5.7 shows the long-run equilibrium of a perfectly competitive firm. The firm appears to be productively efficient because it produces at the lowest point on the *ATC* curve, and it appears to be allocatively efficient because $P = MC$.

SYNOPTIC LINK

In the long run, perfect competition leads to an optimal allocation of good and services which best reflects consumers' preferences. An optimal allocation of resources also implies that there are no externalities in production or consumption. Refer to Chapter 8 for an explanation of externalities.

However, this conclusion must be qualified in a number of significant ways. Productive efficiency combined with allocative efficiency occurs only if:

- All the firms in the market benefit from all the available economies of scale. This is unlikely because each firm produces only a tiny part of total market output. Each firm is likely to be well below minimum efficient scale (*MES*).
- There are perfectly competitive markets for *all* goods and services, including future markets, and $P = MC$ simultaneously in each and every market. To take this point further, *every* firm in *every* market *throughout the world* must be producing where $P = MC$. This is an impossible outcome to achieve, even if all the conditions of perfect competition could be met.
- There are no externalities, negative or positive.

EXTENSION MATERIAL

How competitive is perfect competition?

Although perfect competition is an abstract and unreal market structure, it is interesting to consider the forms that competition might take in a perfectly competitive market economy. The first point to note is that price competition, in the form of price wars or price cutting by individual firms, would not take place. In perfect competition, all firms are passive price-takers, able to sell all the output they produce at the ruling market price determined in the market as a whole. In this situation, firms cannot gain sales or market share by price cutting. Other forms of competition, involving the use of advertising, packaging, brand-imaging or the provision of after-sales service to differentiate a firm's product from those of its competitors, simply destroy the conditions of perfect competition. These are the forms of competition which are prevalent, together with price competition, in the imperfectly competitive markets of the real economy in which we live.

So the only form of competition, both available to firms and also compatible with maintaining the conditions of perfect competition, is cost-cutting competition. Cost-cutting competition is likely in perfect competition because each firm has an incentive to reduce costs in order to make abnormal profit. But even the existence of cost-cutting competition in a perfect market can be questioned. Why should firms finance research into cost-cutting technical progress when they know that other firms have instant access to all market information and that any abnormal profits resulting from successful cost cutting can only be temporary?

Think also of the nature of competition in a perfect market, from the perspective of a typical consumer. The choice is simultaneously very broad and very narrow. The consumer has the doubtful luxury of maximum choice in terms of the number of firms or suppliers from whom to purchase a product. Yet each firm is supplying an identical good or service at exactly the same price. In this sense, there is no choice at all in perfect competition.

A further look at allocative efficiency

As we noted earlier, allocative efficiency occurs when $P = MC$ in all industries and markets in the economy. To explain this further, we must examine closely both P and MC. The price of a good, P, is a measure of the value in consumption placed by buyers on the last unit consumed. P indicates the utility or welfare obtained at the margin in consumption. This is the good's opportunity cost in consumption. For example, a consumer spending £1 on a bar of chocolate cannot spend the pound on other goods. When buying the chocolate bar, other goods or services are sacrificed. At the same time, MC measures the good's opportunity cost in production: that is, the value of the resources which go into the production of the last unit, in their best alternative uses.

Suppose that all the economy's markets divide into two categories: those in which $P > MC$ and those in which $P < MC$. In the markets where $P > MC$, households pay a price for the last unit consumed which is greater than the cost of producing the same unit. The high price discourages consumption, so we conclude that at this price the good is under-produced and under-consumed. Conversely, in the second set of markets, in which $P < MC$, the value (P) placed on the last unit consumed by households is less than the MC of the resources used to produce the last unit. The price is too low, encouraging too much consumption of the good; thus at this price the good is over-produced and over-consumed.

Now suppose resources can be taken from the second group of markets where $P < MC$ and reallocated to the former group of markets in which $P > MC$. Arguably, total consumer welfare or utility will increase as reallocation of resources takes place. As the reallocation proceeds, prices tend to fall in those markets *into which* resources are being shifted and prices tend to increase in the markets *from which* resources are being moved. Eventually, as prices adjust,

KEY TERM
allocative inefficiency occurs when $P > MC$, in which case too little of a good is produced and consumed, and when $P < MC$, in which case too much of a good is produced and consumed.

P equals MC in all markets simultaneously. Beyond the point at which $P = MC$ in all markets, no further reallocation of resources between markets can improve consumer welfare (assuming, of course, that all the other factors which influence welfare, such as the distribution of income, remain unchanged). The outcome, in which $P = MC$ in all markets, is allocatively efficient.

It follows that **allocative inefficiency** occurs when $P > MC$ or $P < MC$. For any given employment of resources and any initial distribution of income and wealth among the population, total consumer welfare can increase if resources are reallocated from markets where $P < MC$ into those where $P > MC$, until allocative efficiency is achieved when $P = MC$ in all markets.

SECTION 5.3 SUMMARY

- It is necessary to understand the six conditions which must be met for a market to be perfectly competitive.
- No real-world market meets all these conditions, but perfect competition provides a yardstick against which the desirable and undesirable properties of real-world markets can be measured.
- In perfect competition short-run equilibrium, firms make abnormal or supernormal profits.
- In perfect competition long-run equilibrium, the entry of new firms has brought the price down until surviving firms make normal profits only.
- Subject to certain conditions, such as a lack of possible economies of scale, in long-run equilibrium, firms are productively efficient and allocatively efficient.

5.4 Monopoly and monopoly power

To remind you again, at this point we are departing from the specification sequence by explaining the topic of 'Monopoly and monopoly power' before we explain the two intermediate market structures of monopolistic competition and oligopoly. While many economists have contributed to the theory of the firm, among the most significant authors is Joan Robinson. Her 1933 book, *The Economics of Imperfect Competition*, was among the first to set out a technical explanation of how firms with some degree of monopoly power might compete with one another.

The meaning of monopoly

KEY TERM
monopoly one firm only in a market.

As we noted earlier in the chapter, pure **monopoly** means one firm only in a market. We quoted regional water companies in the UK as an example. However, the word 'monopoly' is often used in a much looser way to describe any market in which there is a dominant firm, but in which there are also smaller firms. In this and the following sections of this chapter, we will be looking solely at pure monopoly.

SYNOPTIC LINK
Chapter 6 compares *monopsony* or a single *buyer* in the economy's labour markets with *monopoly* or a single *seller* in the economy's goods markets. Both situations involve an organisation having significant power in a market.

Short-run and long-run profit maximisation in monopoly

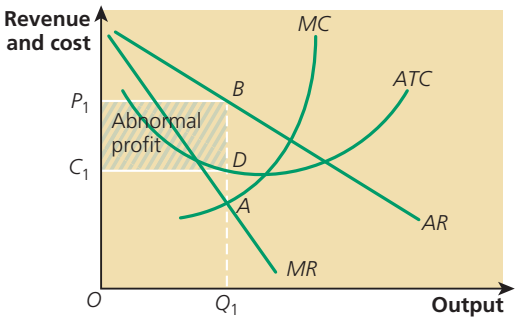


Figure 5.8 Profit maximisation in monopoly

The profit-maximising or equilibrium level of output in monopoly is shown in Figure 5.8. As in perfect competition, the equilibrium output Q_1 is located at point *A*, where *MR* = *MC*. It is worth repeating that *MR* = *MC* is the condition that must be met to maximise profit for any firm, whatever the market structure. However, in monopoly, point *A* does not show the profit-maximising price, which is located at point *B* on the demand curve or *AR* curve above point *A*. P_1 is the maximum price the monopolist can charge and succeed in selling output Q_1.

You will notice that Figure 5.8 does not distinguish between *short-run* and *long-run* profit maximisation in monopoly. This is because a monopoly is protected by barriers to entry, which prevent new firms entering the market to share in the abnormal profit made by the monopolist. Entry barriers enable the monopolist to preserve abnormal profits in the long run as well as in the short run. By contrast, in perfect competition abnormal profits are temporary, being restricted to the short run. Indeed, in monopoly, abnormal profit is often called *monopoly profit*. A monopolist has the **monopoly power** to preserve profit by keeping competitors out.

KEY TERMS

monopoly power (also known as **market power**) the ability of a monopoly to raise and maintain price above the level that would prevail under perfect competition. Market power can also be exercised, usually to a lesser degree, by firms in oligopoly and monopolistic competition.

APPLICATION OF ECONOMICS IN THE REAL WORLD 5.3

Patents and drug companies

Without patents, drug companies would not survive in the market. When rivals try to breach those patents, multinational drug companies react ferociously. They bring in lobbyists. They bring in lawyers. Often — by their own admission — they bring in private detectives.

'Intellectual property protection is critical to an industry like ours,' said the chairman of the UK's largest drug company, GlaxoSmithKline. 'If someone can come along the next day and copy [your invention], then no one would ever put any money into R&D.' He added, 'If we believe people are undermining our intellectual property we have to deal with it by getting the information. Of course we — and I'm sure all other companies — use professional people who work, within the law, to get that information.'

Pfizer also admits to using private detectives: 'They know the loops in the law.' The 'loops' are many. Private detectives raid rubbish bins, track down former employees and prepare 'sting operations' by luring adversaries into doing business with fake companies.

In developed economies, the multinational drug companies are protected by strong patent laws. But in India, China and South America, where there is not currently the same infrastructure to enforce property rights, the task facing drug companies in protecting their patents is far harder.

Follow-up questions

1. What is a patent?
2. Name two other types of intellectual property right in addition to patents.
3. Governments in developing economies often claim that their citizens cannot afford the prices charged for medicines by companies such as GlaxoSmithKline and Pfizer. Do you agree that drug prices should be lower in these places? Justify your answer.
4. Describe how technical progress has been eroding the power of music industry companies to protect their intellectual property.

SECTION 5.4 SUMMARY

- In monopoly, entry barriers prevent the entry of new firms from competing away abnormal profit. Abnormal profit exists in monopoly in the long run as well as in the short run.
- As in all market structures, profits are maximised at the level of output at which *MR* = *MC*.
- A monopoly faces a downward-sloping demand curve (the market demand curve), which is the monopoly's average revenue (*AR*) curve.
- The monopoly's marginal revenue (*MR*) curve is positioned below the *AR* curve.
- Monopoly must not be confused with monopoly power. Monopoly firms possess the power to restrict output and raise the price, but other imperfectly competitive firms also possess a degree of monopoly power.

5.5 The advantages and disadvantages of monopoly

Possible advantages of monopoly

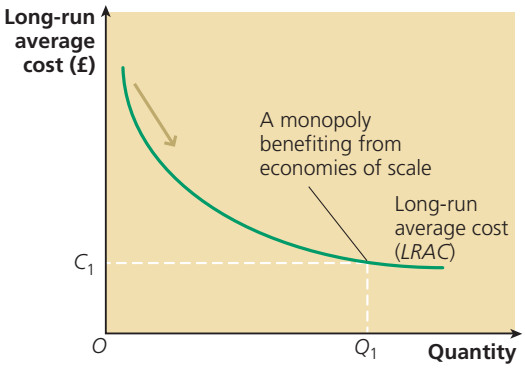


Figure 5.9 How economies of scale may justify monopoly

The possible advantages of monopoly result from two sources: economies of scale and dynamic efficiency. When substantial economies of scale are possible in an industry, monopoly may lead to a better outcome than competition. Figure 5.9 illustrates a natural monopoly, described in the extension material below. Because of limited market size, there is insufficient room in the market for more than one firm benefiting from full economies of scale.

In Figure 5.9, economies of scale are shown by the downward-sloping long-run average cost curve. We assume that a monopoly is able to produce output Q_1 at a long-run average cost (or unit cost) of C_1, whereas competitive firms are unable to produce this output without destroying the competitive market.

A further advantage of monopoly, which we explain in section 5.11, is that compared to perfect competition a monopoly can use its abnormal profit to fund research and development (R&D), which then leads to better ways of making existing products and to the development of completely new products.

TEST YOURSELF 5.8

Explain the main efficiency concepts that economists use when analysing and evaluating different market structures, such as perfect competition and monopoly.

EXTENSION MATERIAL

Natural monopoly

A pure natural monopoly can occur when, for climatic or geological reasons, a particular country or location is the only source of supply of a raw material or foodstuff. However, we usually use the term 'natural monopoly' to describe a situation in which there is only room in the market for one firm benefiting to the full from economies of scale. In the past, utility industries such as water, gas, electricity and the telephone industries were regarded as natural monopolies.

Because of the nature of their product, utility industries experience a particular marketing problem. The industries produce a service that is delivered through a distribution network or grid of pipes and cables into millions of separate businesses and homes. Competition in the provision of distribution grids is extremely wasteful, since it requires the duplication of fixed capacity, therefore causing each supplier to incur unnecessarily high fixed costs.

Possible disadvantages of monopoly

The main disadvantage of monopoly is that it may lead to productive inefficiency and allocative inefficiency. The latter results in resource misallocation. We explained above that a possible advantage of monopoly is that in the long run average costs may fall to be lower than those in perfect competition as a result of a monopoly's ability to benefit from economies of scale and to innovate. The counter-argument is that, protected from competitive pressures, a monopoly may *profit-satisfice* rather than *profit-maximise* and be content with satisfactory profits and an easy life.

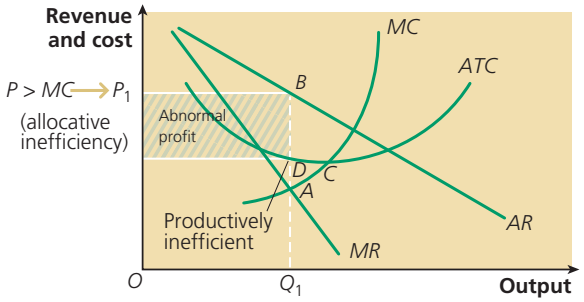


Figure 5.10 Productive and allocative inefficiency in monopoly, assuming an absence of both economies of scale and successful innovation

In the absence of any benefits associated with economies of scale and successful innovation, Figure 5.10 shows that a monopoly is likely to be both productively inefficient and allocatively inefficient. Productive inefficiency occurs because average total costs are shown at point *D* on the diagram and not at point *C*, which would be the case in perfect competition. The profit-maximising level of output Q_1 is also allocatively inefficient because price (shown at point *B*) is greater than marginal cost (shown at point *A*). In summary, for allocative efficiency to occur, price must equal marginal cost ($P = MC$). In monopoly, however, $P > MC$ as Figure 5.10 shows.

EXTENSION MATERIAL

X-efficiency and X-inefficiency

X-efficiency occurs when a firm successfully eliminates all unnecessary costs of production it might otherwise have incurred. X-inefficiency, by contrast, means that unnecessary production costs still persist.

Look back at Figure 5.7, which shows that in long-run equilibrium a perfectly competitive firm is both productively efficient (minimising *ATC*) and allocatively inefficient (producing where $P = MC$). It is also X-efficient, since competitive market forces have eliminated any unnecessary costs of production which in the short run the firm had previously incurred.

Now look at Figure 5.8, which shows the equilibrium level of output in monopoly at Q_1, with the price at P_1. The firm is productively inefficient (average costs are not minimised at point *D*) and it is also allocatively inefficient ($P > MC$). The monopoly is producing too little and selling at too high a price.

However, a monopoly may choose to be X-inefficient rather than X-efficient. To understand this, look closely at Figure 5.11. This diagram is the same as Figure 5.8, but with point *Z* added. If the monopoly produces the level of output Q_1, without incurring unnecessary costs, its average costs per unit of

output are shown at point *D*, which is located on the *ATC* curve. However, if the monopolist produces at point *Z* rather than at point *D*, unnecessary average production costs are incurred, equal per unit to the distance between *Z* and *D*. The monopoly may be able to survive perfectly happily, enjoying an 'easy life' and making *satisfactory* rather than *maximum* profits. This is because entry barriers protect monopolies. As a result, the absence or weakness of competitive forces means unnecessary costs very often persist in monopoly and in other imperfectly competitive markets. In this situation, the monopoly ends up producing 'off' its average cost curve at a point such as *Z*, which means that it is X-inefficient as well as productively and allocatively inefficient.

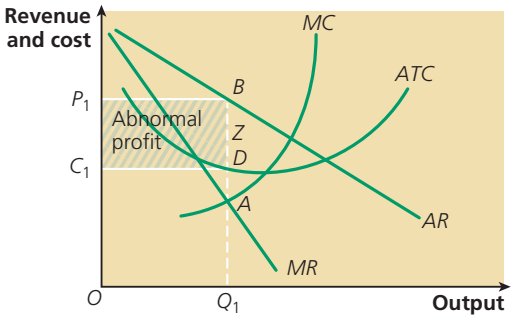


Figure 5.11 X-inefficiency and monopoly

TEST YOURSELF 5.9

Explain the difference between productive efficiency and allocative efficiency.

APPLICATION OF ECONOMICS IN THE REAL WORLD 5.4

Natural monopoly in the UK water industry

In 1989 the UK water industry, which had previously been run largely by publicly owned authorities, was sold to private owners. However, in the 30 or so years since their privatisation, most of the water companies have been bought by foreign owners.

The largest water company, Thames Water, which serves the London region, provides a case in point. Initially, the Australian bank Macquarie borrowed £2.8 billion to buy Thames Water. Then, according to the journalist Nick Cohen writing in *The Spectator* in September 2017, Macquarie loaded £2 billion of Cayman Islands debt on to Thames Water and its customers, despite giving assurances to the water regulator Ofwat that it would do no such thing. This allowed Macquarie to take its profits, leaving behind £2 billion of debt for Thames Water's customers to bear.

Now Thames Water is owned by a Kuwaiti investment fund and a Canadian pension fund. But when pressed by the BBC to say that they would not seek to imitate Macquarie and extract exorbitant returns from a captive market, they refused to answer the question.

In 2017 the *Financial Times* described water privatisation as 'an organised rip-off'. The *FT* posed the rhetorical question: 'How hard can it be to be the chief executive of a privatised British water company?' The paper answered its own question by stating that the company's customers are determined by geography, and its prices set by the industry regulator. Pretty much all the company has to do is to make sure its sewage plants work and to keep the public waterways clear of human waste.

But after loading Thames Water with debt and flooding the Thames Valley with excrement, Thames Water's then chief executive received a 60% pay rise in 2015. At the time of the sell-off of the publicly owned water authorities in 1987, the then-Conservative government argued that greater competition would significantly increase efficiency within the water industry. However, in the 30 or so years since the sell-off, there is very little evidence that efficiency has improved. Many argue the opposite: that efficiency has deteriorated, particularly with regard to the pollution that the water companies discharge into the environment.

Follow-up questions

1. Why is Thames Water a natural monopoly?
2. Why is it difficult, if not impossible, to increase competition in the water industry?
3. Assess the case for taking the water companies back into public ownership.
4. Find out how much the chief executives of water companies are paid. Do you think that they are over-paid? Justify your answer.

QUANTITATIVE SKILLS 5.2

Worked example: profit maximisation and market structure

Firm X is maximising profit in the long run, and the following data are known about its cost and revenue structure:

- **marginal costs = £400**
- **marginal revenue = £400**
- **average costs = £450**
- **average revenue = £500.**

Firm X could be operating in any of the market structures listed below except one. Which is the exception?

Market structures: perfect competition; monopoly; monopolistic competition; oligopoly.

In the data, $MR = MC$, so we know the firm is making the maximum profit. However, $MC < AC$, so the firm is producing on the downward-sloping section of the AC curve. Also, since $MR < AR$, the AR curve slopes downwards. All this is consistent with long-run profit maximisation in monopoly, monopolistic competition and oligopoly, but not with perfect competition, for which the following has to hold: $MR = MC = AC = AR$. Perfect competition is the answer.

SECTION 5.5 SUMMARY

- Monopolies can gain the benefits of economies of scale.
- Economies of scale are shown by a falling long-run average cost curve.
- Monopoly profit can be used to finance R&D with the result that monopolies can be dynamically efficient.
- However, monopolies can abuse their monopoly power, reducing competition, choosing an 'easy life' and incurring unnecessary costs.
- The concepts of productive efficiency and allocative efficiency are used to judge the advantages and disadvantages of monopoly.
- Provided that certain conditions are met, such as a lack of economies of scale, perfect competition is both productively and allocatively efficient, but monopoly is not.

5.6 Monopolistic competition

The meaning of monopolistic competition

In significant respects, monopolistic competition resembles both perfect competition and monopoly.

Monopolistic competition resembles perfect competition in the following ways:

- As in perfect competition, there are a large number of firms in the market.
- In the long run there are no barriers to entry or exit.
- As a result, the entry of new firms, attracted by short-run abnormal profits, brings down the price each firm can charge until only normal profits are made in the long run.

Monopolistic competition also resembles monopoly in the following ways:

- Each firm faces a downward-sloping demand curve. This results from the fact that each firm produces a slightly different product — differentiated by such features of modern production and marketing as style, design, packaging, branding and advertising. The goods produced by the various firms provide *partial* but *not perfect* substitutes for each other. The resulting *product differentiation* in the market means that each firm possesses a degree

of monopoly power over its product. Unlike in perfect competition, if a firm raises its price, it does not lose *all* its customers because there is brand loyalty.

- Each firm's marginal revenue (*MR*) curve is below its average revenue (*AR*) curve, which of course is the demand curve for the firm's output.

STUDY TIP

Students often wrongly believe that monopolistic competition is closer to pure monopoly than it is to perfect competition. In reality, monopolistically competitive markets, characterised by a large number of small to medium-sized firms competing against each other by differentiating the goods or services they produce and sell, are closer to perfect competition.

Short-run profit maximisation in monopolistic competition

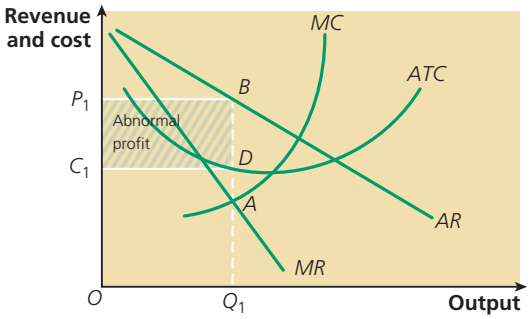


Figure 5.12 Short-run profit maximisation in monopolistic competition

Short-run profit maximisation in monopolistic competition is illustrated in Figure 5.12, which is very similar to profit maximisation in monopoly, illustrated earlier in Figure 5.8. However, in monopolistic competition the demand or average revenue curve represents demand for the goods produced by just *one* firm within the market rather than demand for the output of the *whole* market. And because the other firms within the market produce *partial* though *not perfect* substitutes, the demand curve facing the firm is likely to be rather more elastic at the prices each firm may decide to set than would be the case in pure monopoly.

The profit-maximising level of output, Q_1, is located below point *A* on the diagram, where *MR* = *MC*, and the abnormal profits made by the firm in the short run are shown by the rectangular area C_1DBP_1.

Long-run profit maximisation in monopolistic competition

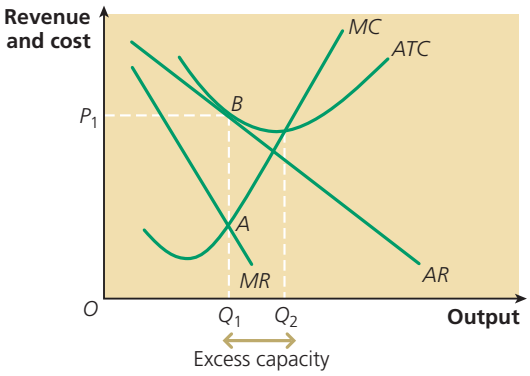


Figure 5.13 Long-run profit maximisation in monopolistic competition

The absence of barriers to entry or exit in the long run is of great importance in the theory of monopolistic competition. Long-run profit maximisation in monopolistic competition is different in an important respect from profit maximisation in monopoly. In the long run, the entry of new firms causes the demand curve or *AR* curve facing an established firm to shift leftwards or inwards. The leftward shift may result from the introduction of new substitute products, attracting some customers away from the existing firms. Long-run profit maximisation is achieved when the *AR* curve just touches (or forms a tangent to) the firm's *ATC* curve, thereby removing the firm's abnormal profit. This is shown in Figure 5.13, at point *B* immediately above level of output Q_1. Since only normal profit is made, total sales revenue and total costs of production are both shown by the rectangle OQ_1BP_1. Note also that the point of tangency between the *AR* and *ATC* curves occurs immediately above point *A*, which determines the profit-maximising level of output Q_1 where *MR* = *MC*.

TEST YOURSELF 5.10

Describe the ways in which monopolistic competition resembles perfect competition and the ways in which it resembles pure monopoly.

STUDY TIP

It is worth remembering that in all market structures — perfect competition, monopoly and all types of imperfect competition — the profit-maximising level of output occurs where *MR* = *MC*.

Evaluating monopolistic competition

We have just explained that, as in perfect competition, the market mechanism eliminates abnormal profits in monopolistic competition and brings about a long-run profit-maximising price which equals the average cost of production ($P = ATC$). Nevertheless, in the absence of any economies of scale, monopolistic competition is both allocatively and productively inefficient in comparison with perfect competition. As in monopoly, $P > MC$ and the firm is producing above the lowest point on its *ATC* curve.

Indeed, because in the long run a firm in monopolistic competition must be producing a level of output less than the output at which average total costs reach their lowest point, the productive inefficiency in monopolistic competition takes the form of excess capacity, measured by the difference in Figure 5.13 between Q_1 (the profit-maximising level of output) and Q_2, the productively efficient level of output, which is located below the lowest point on the *ATC* curve. However, as in perfect competition, long-run profit maximisation and freedom of entry in monopolistic competition mean that firms are forced to produce on their average cost curves. They must eliminate any unnecessary costs; otherwise they will fail to make normal profits.

At the same time, product differentiation increases the range of choice available to the consumer. It can be argued that monopolistic competition can in fact improve economic welfare. The economist Kelvin Lancaster has argued that the number of differentiated products increases until the gain to consumers in choice from adding one more product to the market exactly equals the loss resulting from having to produce fewer of the existing products at a higher cost. According to this argument, monopolistic competition does not necessarily result in economic waste. Consumers may prefer the wider choice available in monopolistic competition to any improvement in productive efficiency that alternative market structures might provide.

TEST YOURSELF 5.11

The price of cotton, a major cost incurred by clothing retailers such as Marks & Spencer, recently increased by 80%. Explain why only one of the five possible outcomes listed on the right is likely to be correct and why the other four are likely to be wrong.

Sales of clothing	Profits
A Constant	Fall
B Rise	Rise
C Fall	Constant
D Constant	Constant
E Fall	Fall

QUANTITATIVE SKILLS 5.3

Worked example: profit maximisation in monopolistic competition

Table 5.1 provides information about the price, total revenue, total costs and marginal costs for haircuts on a particular day in a barber's shop operating in monopolistic competition.

From the information in the table:

(a) fill in the missing information in the marginal revenue and total costs columns

The answer is shown in Table 5.2 below.

(b) calculate the profit-maximising price of haircuts

Profit maximisation occurs at the level of output 3 haircuts per day, at which *MR* = *MC*. This is when the price of a haircut is £16 and both *MR* and *MC* are £12. (Note that total profit is the same at 2 haircuts per day, with the third haircut having zero effect on total profit.)

Table 5.1 Revenue and costs for a barber's shop

Number of haircuts per day	Price (£)	Total revenue (£)	Marginal revenue (£)	Total costs (£)	Marginal costs (£)
0	22	0		10	–
1	20	20			4
2	18	36			6
3	16	48			12
4	14	56			16
5	12	60			18

Table 5.2 Revenue and costs for a barber's shop: completed version

Number of haircuts per day	Price (£)	Total revenue (£)	Marginal revenue (£)	Total costs (£)	Marginal costs (£)
0	22	0	–	10	–
1	20	20	20	14	4
2	18	36	16	20	6
3	16	48	12	32	12
4	14	56	8	48	16
5	12	60	4	66	18

Is advertising a barrier to market entry or does it promote competition?

Advertising can be divided into informative advertising and persuasive advertising. As the name suggests, *informative advertising* increases competition because it provides consumers and producers with useful information about goods and services which are available to buy, and about the different goods that different firms are producing.

By making the demand curve for a product less price elastic, *persuasive advertising*, by contrast, often reduces competition. In effect, customers become 'captive customers' who are unwilling to buy a cheaper substitute good. Persuasive advertising tries to make people believe that a product is a 'must-have' product. Little information about the good itself is provided: for example, its price. Instead,



People who drink Coca-Cola may be unwilling to drink a cheaper substitute

advertisements focus on how ownership or use of the product will improve the consumer's feelings of self-worth and/or the image portrayed to other people.

Persuasive advertising often goes hand in hand with *saturation advertising*. Monopolies and other large firms use saturation advertising to prevent small firms entering the market. The small firms are unable to enter the industry because they cannot afford the minimum level of advertising and other forms of promotion for their goods which are necessary to persuade retailers to stock their products. The mass advertising, brand imaging and other marketing strategies of large established firms effectively crowd out newcomers from the marketplace. Supermarkets are often unwilling to stock goods produced by new entrants to the market because their products are insufficiently advertised.

Price competition

Price competition takes place when a firm reduces prices in order to sell more of a good or service. Increased sales can occur in two different but interrelated ways. In the first, consumers switch from other markets where prices are higher and buy this good instead. In the second, consumers switch from buying similar goods from rival firms *within the same market* to buy the good from the firm that has cut its price.

It is sometimes argued, though without much evidence, that competitive firms do not like to use price competition because it leads to self-defeating price wars, which ultimately only benefit consumers. However, there are a number of ways in which firms undertake price competition, particularly in markets dominated by just a few firms. Besides limit pricing and predatory pricing, which we have already described, firms, especially those in monopolistic competition and competitive oligopoly, may undertake special offer pricing. Firms introduce temporary 'special offer' prices on some of the goods they are selling. Supermarket pricing provides many examples of this, including offering certain goods for sale at discounted prices and multi-buy offers, for a limited period of time.

TEST YOURSELF 5.12

With the help of a diagram, explain why a firm making normal profit in monopolistic competition is likely to be productively inefficient and also allocatively inefficient.

SYNOPTIC LINK

The coverage of price competition in this chapter links with earlier analysis of the role of prices in a market economy in Chapter 3.

Non-price competition

In the absence of, or in addition to, price competition — through which firms try to gain market share and/or protect their existing sales by cutting their prices — firms in imperfectly competitive markets, both in monopolistic competition and in oligopoly, are likely to undertake various forms of *non-price competition*. These include:

- marketing competition, including obtaining exclusive outlets such as tied public houses and petrol stations through which breweries and oil companies sell their products
- the use of persuasive advertising, product differentiation, brand imaging, packaging, fashion, style and design
- quality competition, including the provision of point-of-sale service and after-sales service.

SECTION 5.6 SUMMARY

- Monopolistic competition is 'imperfect competition among the many'.
- Monopolistic competition resembles perfect competition in that abnormal profits can be made in the short run but not in the long run.
- In the long run, the entry of new firms competes away abnormal profits.
- Monopolistic competition resembles monopoly in that each firm faces a downward-sloping demand curve (and *AR* curve).
- In the long run, firms in monopolistic competition are allocatively efficient, but not productively efficient.
- However, consumers benefit from the choice offered by a large number of firms selling slightly differentiated products.

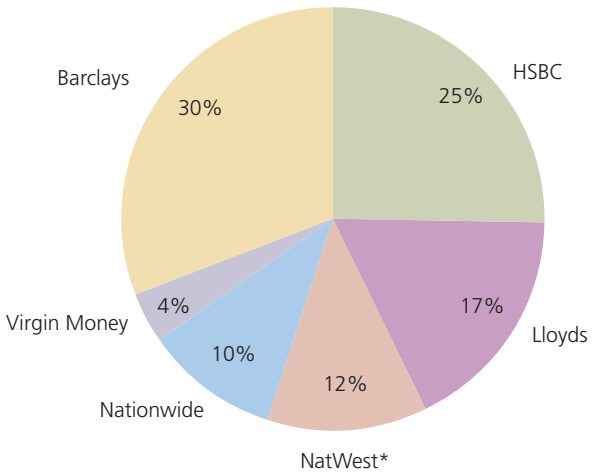
5.7 Oligopoly

The meaning of oligopoly

The prefix 'oligo-' attached to a word is taken from the Greek word *oligos*, meaning few. In a literal sense, therefore, an oligopoly is a market or industry containing a few firms. This is the way oligopoly is defined as a form of market structure. However, there is a problem with this method of definition. How many or how few firms should there be for oligopoly to exist? Do 10 firms make up an oligopoly but 11 firms do not? And what about a market in which there are, say, seven very large firms, but also scores or even hundreds of small firms?

Market structure and concentration ratios

As we have just noted, oligopoly can be defined in terms of market structure and a limited number of firms in a market. We will look at how concentration ratios are used to identify oligopoly in highly concentrated market structures.



*NatWest is a subsidiary of the Royal Bank of Scotland Group.

Note: figures do not total 100% because of rounding.

Source: Statista.com

Figure 5.14 Market shares in the UK banking industry, 2021 (% of total bank assets)

A **concentration ratio** can provide a good indicator of an oligopolistic market structure. For example, a five-firm concentration ratio shows the percentage or share of output in an industry produced by the five largest firms in the industry. Figure 5.14 enables us to calculate the five-firm concentration ratio among the six largest banks in the UK banking industry reported in 2021. The five-firm concentration ratio was 94%, though we should note that many other smaller UK banks such as Metro Bank and the Co-operative Bank are not included in the data.

KEY TERM

concentration ratio measures the market share (percentage of the total market) of the biggest firms in the market. For example, a five-firm concentration ratio measures the aggregate market share of the largest five firms.

APPLICATION OF ECONOMICS IN THE REAL WORLD 5.5

Competition in the UK supermarket industry

Writing in an October 2014 edition of *Management Today*, Alastair Dryburgh explained why, in his view, Tesco cannot compete with Aldi and Lidl. The management expert argued that if Tesco wanted to return to health, it had to give up the idea that it could win by being the cheapest. Tesco was losing out to 'hard discounter' retailers like Aldi and Lidl.

Some commentators were arguing that Tesco should cut prices by more than it had done in the past in order to compete, but Dryburgh believed that this would be an extremely damaging move. Aldi had stated publicly that its intention was to remain at least 15% cheaper than Tesco. Aldi has always been well placed to meet this target: its whole business has been designed to be lower cost than the competition.

In February 2017 Aldi stocked only around 1,500 lines, as against Tesco's 90,000. Around 90% of those lines were reported to be own brands. The package sizes are designed to make optimum use of the shelves, and the goods are delivered in shelf-ready packaging. Dryburgh argued that a company cannot compete on price with other firms which have systematically designed their businesses to be lower cost — it would be like British Airways deciding to compete on price with Ryanair.

Follow-up questions

1. Suggest two forms of non-price competition that Tesco might use to regain market share from 'hard discounter' retailers.
2. Tesco's largest supermarkets have usually been built outside of town and city centres. However, along with other supermarket companies, since 1994 Tesco has invested in convenience stores such as Tesco Express in high-street locations and in suburban shopping parades. Why have supermarket companies such as Tesco chosen to invest in convenience stores?
3. Research the prices charged in Tesco superstores and convenience stores. Are the prices charged similar or different? Suggest one reason for any price differences.
4. Research Tesco's recent business strategy and comment on its success or failure.

QUANTITATIVE SKILLS 5.4

Worked example: calculating a concentration ratio

Table 5.3 shows the usage-based market share of internet web browsers in Europe in July 2022.

Table 5.3 Market share of internet web browsers

Web browser market	Market share (%)
Google Chrome	58.8
Apple Safari	19.9
Mozilla Firefox	6.5
Microsoft Edge	5.9
Samsung Internet	3.5
Others	5.4

Calculate the four-firm concentration ratio for the web browser market in Europe and comment on the nature of the market structure.

We calculate the four-firm concentration ratio by adding up the percentage market shares of the four leading firms. This is 58.8% + 19.9% + 6.5% + 5.9%, which equals 91.1%. The high concentration ratio tells us that the European web browser market is an oligopoly, at least as defined by market structure. The data provide no evidence as to whether or not it is a competitive or collusive oligopoly, though in the absence of evidence discovered by monopoly regulators, such as the Competition and Markets Authority in the UK, it is probably a competitive oligopoly.

TEST YOURSELF 5.13

Distinguish between monopolistic competition and oligopoly.

Oligopoly and market behaviour

KEY TERM

market conduct the pricing and marketing policies pursued by firms. This is also known as market behaviour, but is not to be confused with market performance, which refers to the end results of these policies.

As already mentioned, oligopoly is best defined, not only by market structure or the number of firms in the market, but also by **market conduct**, or the behaviour of the firms within the market. An oligopolistic firm affects its rivals through its price and output decisions, but its own profit can also be affected by how rivals behave and react to the firm's decisions. Suppose, for example, the firm reduces its price in order to increase market share and boost profit. Whether the price reduction increases or reduces the firm's profit depends on the reactions of the other firms.

Interdependence and uncertainty in oligopoly

Competitive oligopoly exists when the rival firms are interdependent in the sense that they must take account of each other's reactions when forming a market strategy, but independent in the sense that they decide their market strategies without cooperation or collusion. As a result, uncertainty is a characteristic of competitive oligopoly; a firm can never be completely certain of how rivals will react to its price, marketing and output strategy. If the firm raises its price, will the rivals follow suit or will they hold their prices steady in the hope of gaining sales and market share?

TEST YOURSELF 5.14

What is the difference between market structure and market conduct?

Non-collusive oligopoly, collusive oligopoly and cartels

As we noted in the previous paragraph, in competitive oligopoly, firms act independently in the sense that they do not form agreements with each other. This is also known as non-collusive oligopoly. The uncertainty facing competitive oligopolists can be reduced and perhaps eliminated by the rival firms colluding together — for example, by forming a **cartel** or price ring. In Figure 5.15, five firms jointly agree to charge a price to keep Firm E, which is the least efficient firm, in the market. In a competitive market, Firm E would have to reduce costs or go out of business. Cartel agreements enable inefficient firms to stay in business, while other more efficient members of the price ring enjoy abnormal profit. By protecting the inefficient and enabling firms to enjoy an easy life protected from competition, cartels display the disadvantages of monopoly (high prices and restriction of choice). However, this is usually without the benefits that monopoly can sometimes bring, namely economies of scale and improvements in dynamic efficiency.

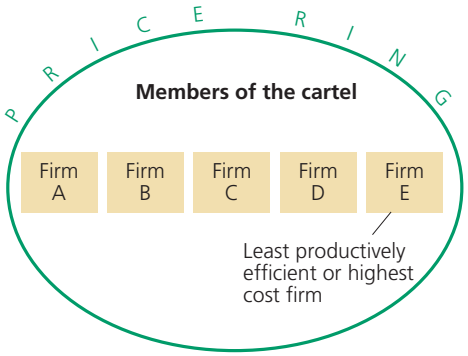


Figure 5.15 A cartel or price ring

TEST YOURSELF 5.15

Why may firms in highly concentrated markets decide to form a cartel?

KEY TERM

cartel a collusive agreement by firms, usually to fix prices. Sometimes there is also an agreement to restrict output and to deter the entry of new firms.

Collusion versus market cooperation

Although forming a cartel can achieve a better outcome than competitive behaviour for oligopolistic firms, the result is unlikely to be good for consumers. For this reason, cartel agreements are usually illegal and judged by governments as being anti-competitive and against the public interest.

Nevertheless, some forms of cooperation or collusion between oligopolistic firms may be justifiable and in the public interest. These include joint product development (such as the Ford Fiesta and Mazda 2 cars, both built using many shared components between 1974 and 2015), and cooperation to improve health and safety within an industry or to ensure that product and labour standards are maintained. Such examples of industry collaboration, or overt collusion, which is in full public view, are normally deemed to be good, in contrast to price collusion, which is regarded as bad. Price collusion and other market-rigging agreements almost always take place in secret. This is covert collusion. Tacit collusion, by contrast, occurs when there is 'an understanding' without any explicit agreement between the firms.

STUDY TIP

Collusive or cooperative behaviour enables firms to reduce the uncertainty they face in imperfectly competitive markets. However, some forms of collusion — for example, on joint development of products or ensuring industry safety standards — are in the public interest.

TEST YOURSELF 5.16

The chairpersons of firms tendering for a government contract attend a secret meeting and verbally agree with each other to fix a higher price than if there were independent bids. What sort of collusion is this? Justify your answer.

EXTENSION MATERIAL

Joint-profit maximisation in collusive oligopoly

Joint-profit maximisation, which is illustrated in Figure 5.16, occurs when a number of firms decide to act as a single monopolist, yet keep their separate identities. Oligopolistic firms undertake joint-profit maximisation in the belief that it can lead to higher profits for all the firms taking part. The monopoly *MC* curve depicted on the right-hand side of the diagram is the sum of the identical *MC* curves of three firms (one of which is shown on the left of the diagram). The three firms share an output of 750 units, determined on the right of the diagram where the industry *MR* and *MC* curves intersect. Each firm charges a price of £10, which, as the diagram shows, is the maximum price consumers are prepared to pay for 750 units of the good. The monopoly output of 750 units is well below 1,000 units, which would be the output if the industry were perfectly competitive. The shaded area in the right-hand panel shows the efficiency or welfare loss caused by the cartel raising the price to £10 and restricting output to 750 units. In this example, the members of the cartel split the 750 units equally, each firm producing 250 units. The shaded area on the left of the diagram shows the abnormal profit made by each firm.

Although there is an incentive for firms to collude to maximise their joint profits, there is also an incentive for each member of the cartel to cheat on the agreement. The marginal cost of producing the 250th unit of the good is only £4, yet for the firm (but not the whole industry) the marginal revenue received from selling one more unit is £10 (i.e. the price set by the cartel). One member of the cartel can increase its profit at the expense of the other firms by secretly selling an output over and above its quota of 250 units at a price less than £10, but greater than the marginal cost incurred (£4). This is an example of a divergence between individual and collective interest. The firms' collective interest is to maintain the cartel so as to keep total sales down and the price up. But each firm can benefit by cheating on the agreement — as long as all the others do not also cheat.

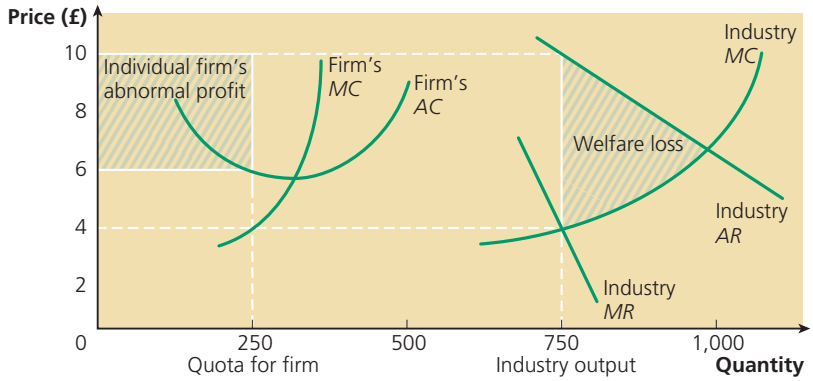


Figure 5.16 Joint-profit maximisation by members of a cartel

SYNOPTIC LINK

Economic welfare is a key measure of how successful a firm or economy is in achieving an efficient allocation of scarce resources. Welfare gains and welfare losses are explained in relation to international trade on page 533.

The kinked demand curve model

The kinked demand curve theory can be used to illustrate how a competitive oligopolist may be affected by rivals' reaction to its price and output decisions. The theory was originally developed to explain alleged price rigidity and an absence of price wars in oligopolistic markets.

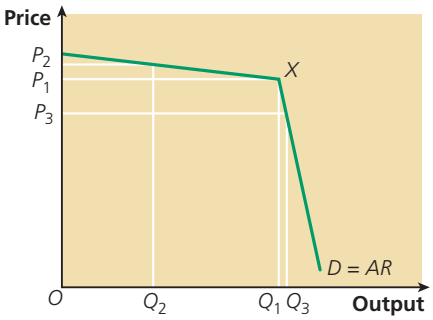


Figure 5.17 The kinked demand curve

Suppose an oligopolist initially produces output Q_1 in Figure 5.17, selling this output at price P_1. In order to anticipate how sales might change following a price change, firms need to know the position and shape of the demand and revenue curves for their products. But in imperfectly competitive markets, firms lack accurate information about these curves, particularly at outputs significantly different from those currently being produced. This means that the demand curve or *AR* curve in Figure 5.17 is not necessarily the correct or *actual* demand curve for the oligopolist's output. Instead, it represents the firm's *estimate* of how demand changes when the firm changes the price it is charging.

When *increasing* price from P_1 to P_2, the oligopolist expects rivals to react by keeping their own prices stable and not following suit. By holding their prices steady, rivals try to gain profit and market share at the firm's expense. This means that the oligopolist expects demand to be *relatively elastic* in response to a price increase. The rise in price from P_1 to P_2 is likely to result in a *more than proportionate fall in demand* from Q_1 to Q_2.

Conversely, when *cutting* its price from P_1 to P_3, the oligopolist expects rivals to react in a very different way, namely by following suit immediately with a matching price cut. In this situation, because the market demand curve for the products of all the firms slopes downwards, each firm will benefit from *some* increase in demand. However, the oligopolist fails to gain sales from rivals *within* the market. This means the oligopolist expects demand to be less elastic, and probably *inelastic*, in response to a price cut. The fall in price from P_1 to P_3 may result in a *less than proportionate increase in demand* from Q_1 to Q_3. The oligopolist therefore expects rivals to react *asymmetrically* when price is raised or lowered.

In Figure 5.17, the oligopolist's initial price and output of P_1 and Q_1 intersect at point *X*, or at the kink at the junction of two demand curves of different elasticity, each reflecting a different assumption about how rivals may react to a change in price. If when price is raised, demand is elastic, and when price is cut, demand is inelastic, any change in price will reduce the oligopolist's total revenue. In this situation, the oligopolist fears that both a price increase and a price cut are likely to reduce total profit. Given this fear, the best policy may be to leave price unchanged.

EXTENSION MATERIAL

Developing the kinked demand curve theory

By developing the theory a little more, we can explain a second reason why prices may tend to be stable in oligopoly. As Figure 5.18 illustrates, a mathematical discontinuity exists along the vertical line drawn above output Q_1. For the demand and average revenue curves, the discontinuity occurs at the 'kink' where the curves intersect. But for the marginal revenue curves, which are twice as steep as the *AR* curves with which they are associated, the discontinuity is the 'gap' between the two *MR* curves, shown by the distance *B* to *C*.

Suppose initially the firm's marginal cost curve is MC_1, intersecting the *MR* curve at point *A*, which is positioned in the middle of the vertical section. The diagram shows that the *MC* curve can rise or fall within the vertical section of the *MR* curve, without altering the profit-maximising output Q_1 or price P_1. But if marginal costs rise above MC_2 at point *B* or fall below MC_3 at point *C*, the profit-maximising output changes. In either of these circumstances, the oligopolist would have to set a different price to maximise profits, provided of course that the *AR* curve accurately measures demand for the firm's product at different prices. Nevertheless, the oligopolist's selling price remains stable at P_1 as long as the marginal cost curve lies between MC_2 and MC_3, despite quite considerable changes in marginal costs.

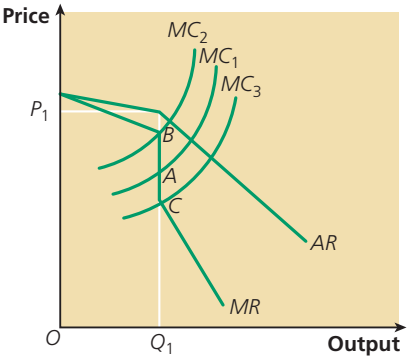


Figure 5.18 Developing the theory of the kinked demand curve

Criticisms of the kinked demand curve theory

There are a number of weaknesses in the theory we have just described. Although at first sight it is attractive as a neat and apparently plausible explanation of price stability in conditions of oligopoly, few economists now accept the kinked demand theory of oligopoly pricing.

First, it is an incomplete theory, since it does not explain how and why a firm chooses in the first place to be at point *X* in Figure 5.17. Second, the evidence provided by the pricing decisions of real-world firms gives little support to the theory. Competitive oligopolists seldom respond to price changes in the manner assumed in the kinked demand curve theory. It is more reasonable to expect a firm to test the market — that is, to raise or lower its selling price to see if rivals react in the manner expected. If rivals do not, then the firm must surely revise its estimate of the shape of the demand curve facing it. Research has shown fairly conclusively that oligopoly prices tend to be stable or sticky when demand conditions change in a predictable or cyclical way, but that oligopolists usually raise or lower prices quickly and by significant amounts, both when production costs change substantially, and when unexpected shifts in demand occur.

STUDY TIP

Students often wrongly believe that the kinked demand curve provides a complete theory of oligopoly. It is actually a very doubtful theory, but it does illustrate how oligopolists are interdependent and affected by uncertainty.

The advantages and disadvantages of oligopoly

The term 'oligopoly' covers a range of more narrowly defined market structures, ranging from pure duopoly (two firms only in a market), to markets which combine quite a large number of firms, a few dominant, but many very

small and without much market power. It also ranges from competitive or non-collusive oligopoly to cartels in which member firms collude together, often to reduce competition. The following list of advantages and disadvantages is not relevant for all oligopoly market structures.

Possible advantages of oligopoly

- Just like a monopoly, firms benefit from economies of scale in oligopoly, which means they can become more dynamically efficient and can pass on cost cuts as low prices to consumers.
- With only a few firms available from which to buy, it will be easy for consumers to compare and choose the best option for their needs. Oher markets may offer too much choice, leading to confusion among consumers. Behavioural economists sometimes make this point.
- Provided that there is a degree of competition, oligopolists continuously innovate and develop new and better products.

Possible disadvantages of oligopoly

- Again, just like a monopoly, oligopolies restrict output and raise prices (and profit), compared to a more competitive market. Firms may satisfice, content with an easy life.
- As we have noted, cartels are a bad form of market structure, combining the disadvantages of monopoly (high prices, productive and allocative inefficiency and lack of choice) with few if any of the benefits.
- Small, competitive and innovative firms may find it difficult to enter the market.
- Ultimately, as in monopoly, the producer rather than the consumer ends up being dominant, with producer sovereignty rather than consumer sovereignty ruling the market.

Price leadership, price agreements and price wars

Price leadership

Because overt collusive agreements to fix the market price, such as cartel agreements, are usually illegal, imperfectly competitive firms often use less formal or tacit ways to coordinate their pricing decisions. An example of covert collusion is **price leadership**, which occurs when one firm becomes the market leader and other firms in the industry follow its pricing example.

KEY TERMS

price leadership the setting of prices in a market, usually by a dominant firm, which is then followed by other firms in the same market.

price agreement an agreement between a firm, similar firms, suppliers or customers regarding the pricing of a good or service.

price war occurs when rival firms continuously lower prices to undercut each other.

Price agreements

We have already seen earlier in the chapter how members of a cartel often fix the prices that all the members of the cartel charge by forming a price ring. **Price agreements** can also be made between firms and their suppliers, and between firms and their customers. In both these cases, a price agreement is usually good for a specified period of time, such as six months.

Price wars

Price wars, which take place both in monopolistic competition and in oligopoly, may be started accidently or may be instigated deliberately to damage competitors. Whereas price leadership usually involves quite friendly relations between the companies involved, price wars — as the name indicates — centre on price cutting aimed at the very least at increasing market share, and in

the extreme at forcing rival firms out of business. Consumers may of course benefit from a price war, at least in the short run, though if firms are driven out of the market, the monopoly power of the surviving firms increases, which is likely to be to the detriment of consumers.

APPLICATION OF ECONOMICS IN THE REAL WORLD 5.6

Discounting in the books market



Ending retail price maintenance in the UK books market resulted in the closure of many small, independent bookshops

Many years ago, UK law allowed manufacturers to decide the prices at which retailers sold their goods. If a shop tried to undercut or discount a set price, the manufacturer could stop supplying the good to the retailer. As a result, price competition between firms selling similar goods hardly existed and manufacturers made excessive profits. This was a restrictive practice known as *retail price maintenance*.

Eventually the law was changed in an effort to promote more competition. However, manufacturers then set recommended retail prices (RRPs), which they hoped shops would abide by.

Selling at recommended retail prices became especially important in book publishing, where very little price discounting took place. Paperback books were generally sold at the prices printed by the publishers on the covers. Small bookshops justified the lack of price competition on the grounds that consumers benefited, both from bookshops surviving in small towns and from a much wider selection of books being on display.

However, the law was eventually changed. Large book chain stores such as Waterstones began to sell books at prices well below the publishers' recommended prices. Small bookshops could not compete and many closed down. However, large stores themselves were now facing increasing competition from the growth of the online retailer, Amazon. Many people now pop into bookshops, not to buy books, but to browse, before going home to order the books they want online.

The picture is different in France. French law fixes book prices with the result that readers pay the same whether they buy online, from a big high-street chain, or from a small bookseller. Extensive discounting is banned, although 5% discounts are allowed. Result: there are between 2,500 and 3,000 independent bookshops in France, compared with fewer than 1,000 in the UK. Most small French towns have at least two bookshops and there is a wide choice of books on display.

The French government says that the banning of discounts of more than 5% has saved its independent bookstores from the free-market capitalism that hit the UK when it abandoned fixed prices. Nevertheless, the owners of French bookshops still argue they cannot compete with Amazon, even with Amazon's discounts limited to 5%, because the online retailer provides free postage and free fast delivery deals on top of the discount. Consumers can also bypass French law by ordering books online in countries such as Belgium.

A French culture minister has said: 'Everyone has had enough of Amazon, which by dumping practices, slashes prices to get a foothold in markets, only to raise them as soon as they have established a virtual monopoly ... the book and reading sector is facing competition from certain sites using every possible means to enter the French and European book market ... it is destroying bookshops.'

Follow-up questions

1. Explain how the changing nature of competition in the book market illustrates the process of creative destruction.
2. Do you agree that the growth of Amazon has been good for consumers? Justify your answer.
3. Conduct some research to find out whether the number of specialist bookshops has declined in the town or area in which you live.
4. Many people are worried about the decline of 'high street' shopping in the UK. Do you agree that this is a problem? What, if anything, can be done to reverse this decline?

SECTION 5.7 SUMMARY

- Oligopoly is imperfect competition 'among the few'.
- High concentration ratios are often used to define oligopoly.
- It is usually more useful to define oligopoly by interdependence and reactive behaviour.
- It is also useful to distinguish between collusive and competitive oligopoly.
- Oligopolists sometimes collude to reduce uncertainty and to increase monopoly profit.
- The theory of the kinked demand curve, which is often used to model competitive oligopoly, illustrates the effects of uncertainty and interdependence in oligopoly, but the theory has a number of weaknesses.
- Just like a monopoly, firms benefit from economies of scale in oligopoly, which means they can become more dynamically efficient and can pass on cost cuts as low prices to consumers.
- Provided that there is a degree of competition, oligopolists continuously innovate and develop new and better products.
- Again, just like a monopoly, oligopolies restrict output and raise prices (and profit), compared to a more competitive market. Firms may satisfice, content with an easy life.
- Collusive oligopoly in the form of a cartel is a bad form of market structure, combining the disadvantages of monopoly (high prices, productive and allocative inefficiency and lack of choice) with few if any of the benefits.

5.8 Price discrimination

The meaning of price discrimination

KEY TERM
price discrimination charging different prices to different customers for the same product or service, with the prices based on different willingness to pay.

Price discrimination involves firms charging different prices to different customers for the same product or service, based on differences in the customers' ability and willingness to pay. Those customers who are prepared to pay more are charged a higher price than those who are only willing to pay a lower price. In the main form of price discrimination, the different prices charged are not based on any differences in costs of production or supply. However, in one form of price discrimination, bulk buying, consumers buying larger quantities are charged lower prices than consumers purchasing smaller quantities of the good.

When this happens, different costs of supply may be involved. Bulk purchases generally have lower average costs of production than smaller purchases.

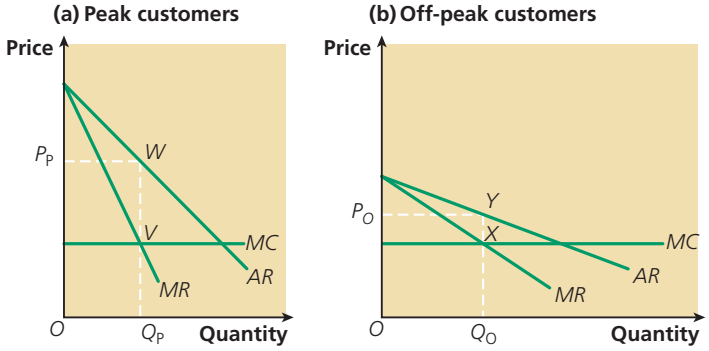


Figure 5.19 Price discrimination when a firm charges different prices to two groups of customers

In Figure 5.19 a rail operating firm divides its market into peak time (busy periods) and off-peak time (less busy periods) customers, each with a different elasticity of demand for travel. For both peak and off-peak customers, the downward-sloping demand curves in Figure 5.19 show average revenue, but not marginal revenue. In each case, the *MR* curve is twice as steep as the *AR* curve. The diagrams also assume that the marginal cost incurred when an extra customer uses the rail service is always the same. This is shown by the horizontal *MC* curve.

To maximise profit, *MR* must equal *MC* in both peak and off-peak sub-markets. As the diagrams show, this means that peak customers pay a higher ticket price than off-peak customers, namely P_P, with off-peak customers paying the lower entry price of P_O. With the different prices being charged,

Q_P peak customers and Q_O off-peak customers use the rail service. The point to note is that the different prices charged result from the different price elasticities of demand. Profit is maximised when more price-sensitive off-peak customers pay less to travel than the less price-sensitive peak customers. Note that when profit maximising, the *MR* received from the last peak customer and off-peak customer admitted are the same. If this were not the case, the rail operator could increase profit by changing the numbers of peak and off-peak tickets available.

The conditions necessary for price discrimination

Successful price discrimination requires that the following conditions are met:

- It must be possible to identify different groups of customers or sub-markets for the product. This is possible when customers differ in their knowledge of the market or in their ability to shop around. Some customers may have special needs for a product and competition among oligopolists may vary in different parts of the market. In some geographical areas and for some products, a firm may face many competitors, whereas in other parts of the market the firm may be the sole supplier.
- At any particular price, the different groups of customers must have different price elasticities of demand. In these circumstances, total profits are maximised by charging a higher price in a market in which demand is less elastic.
- The markets must be separated to prevent seepage. Seepage takes place when customers buying at the lower price in one sub-market resell in another sub-market at a price which undercuts the oligopolist's own selling price in that market. In the European car market, car manufacturers have often charged higher prices for a vehicle in the UK market than in mainland Europe. One reason for this has been the fact that UK motorists demand right-hand drive cars, in contrast to cars in mainland Europe which are left-hand drive. However, seepage occurred when specialist car importers bought cars on the Continent to resell in the UK market, thereby undercutting the prices the car manufacturers recommended for the UK market.

SECTION 5.8 SUMMARY

- Price discrimination involves firms charging different prices to different customers for the same product or service, based on differences in the customers' ability and willingness to pay.
- Successful price discrimination requires a number of conditions are met.
- It must be possible to identify different groups of customers or sub-markets for the product.
- At any particular price, the different groups of customers must have different price elasticities of demand.
- The markets must be separated to prevent seepage.

5.9 The dynamics of competition and competitive market processes

Traditionally, economists have thought of competition as price competition that leads to consumers benefiting from lower prices and more choice. The economy also benefits from the weeding out of high-cost, productively inefficient firms.

However, in imperfectly competitive real-world markets, firms do not just compete on the basis of price, but that competition will, for example, also lead firms to strive to improve products, reduce costs and improve the quality of the service provided. As we saw in section 5.6 on monopolistic competition, imperfectly competitive firms undertake various forms of non-price competition, such as marketing competition, product differentiation, and the use of brand imaging, packaging, fashion, style and design competition. All of these are used to make firms more dynamically efficient and attractive to customers. Non-price competition may also enable innovative and entrepreneurial firms to enter markets more easily.

But on the other hand, business decisions which at first sight may appear to be competitive and in consumers' interests are in fact 'anti-competitive' in the sense that they aim to increase the market power of already dominant firms. This view of competition is closely linked to the process of creative destruction, mentioned in Chapter 4 and earlier in this chapter, and also in the extension material that follows. Creative destruction, which is part of the process through which capitalist economies change over time and become more dynamically efficient, has become the centrepiece for modern thinking on how economies evolve.

EXTENSION MATERIAL

Profit and loss in perfect competition

Figure 5.20 shows a firm facing different prices in a perfectly competitive market. If initially the price is P_1, in order to maximise profit, each firm produces output Q_1 which is located below point *A* on the diagram where *MR* = *MC*. At this price, all firms make abnormal profit.

However, abnormal profit attracts new firms into the market, which causes the price to fall. As the price falls, some abnormal profit continues to be made, as long as the ruling price remains above P_2, which is *break-even price*. Below P_2, the price line is below *ATC* and losses are now made.

Provided, however, that the price is lower than P_2 but higher than P_4, firms make losses in the short run, but these losses are smaller than would be the case if the firms produced no output at all.

This is because in the short run, firms incur fixed costs of production. If they produce zero output, each firm's loss equals fixed cost.

By producing the output at which *MR* = *MC*, firms can do better than this. For example, at price P_3 in Figure 5.20 a firm earns enough revenue to more than cover its variable costs of production. As it can now pay off a *part* of its fixed costs, it is better producing and selling *some* output, thereby incurring a loss *smaller* than would be the case if it produced nothing at all. If, however, the price falls *below* P_4, in the long run the firm should sell its capital assets and shut down. By so doing, the firm reduces its loss to zero. Price P_4 is called *shutdown price*.

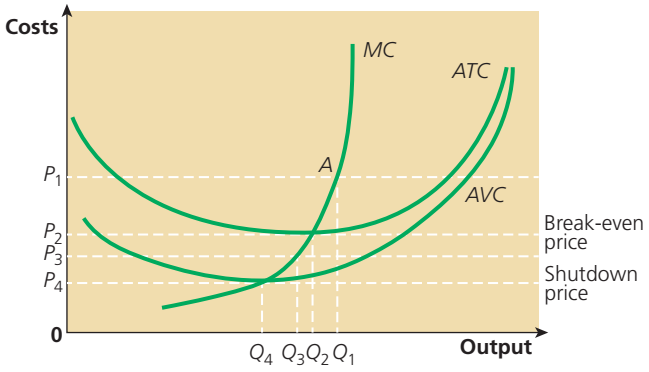


Figure 5.20 Firm facing different prices in a perfectly competitive market

APPLICATION OF ECONOMICS IN THE REAL WORLD 5.7

How Apple competes

Nearly 40 years ago, when the Apple Macintosh personal computer was launched, all rival computers were square boxes and most of them were battleship grey. Not much thought was given to design. But the introduction of the Apple Mac changed things completely.

The first way in which Apple separated itself from the crowd was to 'think different'. Apple's main method of competition is to create new products, and new categories of products, which did not exist before. Examples over the years include innovation in the markets for MP3 and MP4 players, smartphones, tablets and, of course, computers. By not being much bothered about what its competitors do, Apple has broken the mould with all these products and has set industry standards.

Apple also competes through the management style the company has had since its inception. In rival companies, competitive decisions take a very long time to make because they have to be approved by many committees, all of which feed into the decision-making process. Apple is not like this. All decisions are made by a single executive committee. This leads to an almost seamless roll-out of new products, and the ability to launch a new product which nobody else had thought of almost every other year.

The third way in which Apple sets itself aside from its competitors is design. It is said that 'great design principles are pervasive in the Apple DNA'.

Unless Apple's competitors start innovating on their own, Apple will continue to have at least a two-year lead over them. Tim Bajarin, the president of Creative Strategies Inc., believes that thanks to Apple's ability to 'think different', its management style and its design DNA, the company will keep its competitors following it instead of leading the market forward themselves.

Follow-up questions

1. Explain how Apple's success illustrates the dynamics of competition and competitive market processes.
2. Research how Apple has marketed new products in the years between when this book was published (2023) and your reading of the case study.
3. Apple designs computers and other electronic devices, but Apple products are manufactured in China and in other low-wage countries. Former president Donald Trump has argued that such manufacturing should be returned to the USA. Argue the case for and against this.
4. Do you think that Apple will eventually decline due to the process of creative destruction?

SECTION 5.9 SUMMARY

- Both consumers and firms benefit from price competition.
- However, firms in imperfectly competitive market structures also benefit from various forms of non-price competition.
- These may sometimes be 'anti-competitive' in that they increase the market power of dominant firms.
- However, innovative, entrepreneurial firms may also benefit.
- And through the process of creative destruction, dynamic efficiency and economic performance may improve.

5.10 Contestable and non-contestable markets

There is debate about the best way to deal with the problems posed by market concentration and monopoly power, and how to deregulate and remove barriers to market entry. This debate reflects the growing influence of

KEY TERM

contestable market a market in which the potential exists for new firms to enter the market. A perfectly contestable market has no entry or exit barriers and no sunk costs, and both incumbent firms and new entrants have access to the same level of technology.

contestable market theory. Until about 40 years ago, government approaches to the abuse of monopoly power involved an ever-increasing extension of government regulation into the activities of private-sector firms. Increased intervention was justified by the belief that regulatory powers must be strong enough, first, to countervail the growing power of large business organisations and, second, to make monopolies behave in a more competitive fashion.

At the time, monopoly was normally defined by the number of firms in the market and by the share of the leading firms, measured by a concentration ratio. The basic dilemma facing the policy-makers centred on how to reconcile the potential gains of large-scale productive efficiency with the fact that lack of competitive pressure can lead to monopoly abuse and consumer exploitation.

However, in contestable market theory, monopoly power is dependent not on the number of firms in the market or on concentration ratios, but rather on the ease or difficulty with which new firms may enter the market. Industrial concentration is not a problem, provided that an absence of barriers to entry and exit creates the ability for new firms to enter and contest the market. *Actual* competition in a market is not essential. The threat of entry by new firms or *potential* competition is quite enough, according to contestable market theory, to ensure efficient and non-exploitative behaviour by existing firms within the market.

In recent years, contestable market theory has had a major impact on UK monopoly policy. The theory implies that, provided there is adequate potential for competition, a conventional regulatory policy is superfluous. Instead of interfering with firms' pricing and output policies, the government should restrict the role of monopoly policy to discovering which industries and markets are potentially contestable. Deregulatory policies should be used to develop conditions in which barriers to entry and exit are minimised, to ensure that reasonable contestability is possible. Appropriate policies suggested by the theory of contestable markets include removal of:

- licensing regimes for public transport and television and radio transmissions
- controls over ownership, such as exclusive public ownership
- price controls that act as a barrier to entry, such as those which used to be practised in the aviation industry.



A reduction of price controls in the aviation industry has increased competition

Sunk costs and hit-and-run competition

KEY TERM
hit-and-run competition occurs when a new entrant can 'hit' the market, make profits and then 'run', given that there are no or low barriers to exit.

For a market to be perfectly contestable, there must be no barriers to entry or exit and hence no sunk costs. Sunk costs, as described earlier, are costs incurred when entering a market that are irrecoverable should the firm decide to leave the market. *Sunk costs* must not be confused with *fixed costs*, although some sunk costs are also fixed costs. Suppose a firm invests in new machinery when it enters the market. This is a fixed cost, but if the machinery can be sold at a good price to another firm, it is not a sunk cost. In this situation, the cost can be recovered if the firm decides to leave the market. By contrast, if the machinery has no alternative use and a cost of disposal rather than a second-hand value, investment in the fixed capital is a sunk cost. Another sunk cost might be expenditure on advertising to establish the firm in the market. If market entry is unsuccessful and the firm decides to leave, the expenditure cannot be recovered.

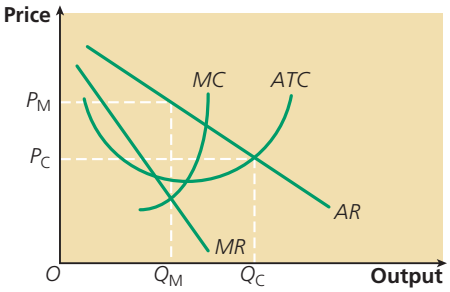


Figure 5.21 A monopoly setting a contestable price P_C to prevent the market being contested

Because sunk costs are unrecoverable and therefore make exit from a market costly, they discourage **hit-and-run competition**. Hit-and-run competition occurs when a firm temporarily enters a market to share in the abnormal profit being made by firms already in the market, and then leaves when abnormal profits have been competed away. As such, it is a feature of a contestable market. And just as *potential* ease of entry makes a market contestable, so the threat of hit-and-run competition can be sufficient to keep prices and profits at their lowest possible levels, thereby increasing consumer surplus. This situation is shown in Figure 5.21, which illustrates how a monopolist facing the threat of potential market entrants reacts by reducing the selling price from P_M to P_C, with the result that only normal profits are made. (This pricing policy is basically the same as limit pricing.)

TEST YOURSELF 5.17

Explain why one of the five possible answers to the following question is correct, but the other four are wrong.

Abnormal profits being made by a perfectly competitive firm in the short run would disappear in the long run because of:

A firms engaging in large-scale advertising
B allocative inefficiencies
C differentiated goods
D freedom of entry into this market
E high sunk costs

SECTION 5.10 SUMMARY

- In contestable market theory, monopoly power is dependent not on the number of firms in the market or on concentration ratios, but rather on the ease or difficulty with which new firms may enter the market.
- In a contestable market, the potential exists for new firms to enter the market.
- A perfectly contestable market has no entry or exit barriers and no sunk costs, and both incumbent firms and new entrants have access to the same level of technology.
- Potential competition is enough to ensure efficient and non-exploitative behaviour by existing firms within the market.
- Contestable market theory now has a major impact on UK monopoly policy.
- For a market to be perfectly contestable, there must be no sunk costs.
- Hit-and-run competition occurs when a new firm enters the market, makes profits and then quickly leaves, given that there are no or low barriers to exit.

5.11 Market structure, static efficiency, dynamic efficiency and resource allocation

KEY TERM
static efficiency efficiency (e.g. productive and allocative efficiency) at a particular point in time.

With the exception of the explanation of static efficiency which follows, this section provides a summary of key concepts explained earlier in the chapter, and on occasion in Chapters 1 and 4.

So far in this chapter we have examined different market structures in terms of whether or not they are productively and allocatively efficient. In Chapter 1 we defined productive efficiency as the minimisation of average costs and earlier in this chapter we defined allocative efficiency as occurring when price equals marginal costs ($P = MC$). These types of efficiency can be thought of as **static efficiency**, which is efficiency measured at a particular point of time.

Monopolies and imperfectly competitive firms can also benefit from dynamic efficiency, which is improvements in other types of efficiency, particularly productive efficiency, that occur over time, and which result from technical progress and innovation. We defined dynamic efficiency in Chapter 4 on page 114. Monopolies can make abnormal profit, which in this context is also called monopoly profit, in the long run as well as in the short run. This profit can achieve improvements in dynamic efficiency through funding research and development (R&D) and product innovation.

Improvements in dynamic efficiency change the allocation of resources within the economy. Resources shift away from high-cost industries into lower-cost industries, and into better ways of making existing products and the development of completely new products.

SECTION 5.11 SUMMARY

- Static efficiency is productive and allocative efficiency measured at a particular point of time.
- Earlier sections of the chapter explained that in long-run equilibrium, perfect competition is both productively and allocatively efficient.
- Monopolistic competition is allocatively efficient in the long run, but not productively efficient.
- Dynamic efficiency is improvements in productive efficiency occurring over time.
- Monopoly and oligopoly are likely to be dynamically efficient.

5.12 Consumer and producer surplus

KEY TERM
consumer surplus a measure of the economic welfare enjoyed by consumers: surplus utility received over and above the price paid for a good.

In this final section of the chapter, we use the concepts of consumer surplus and producer surplus as welfare criteria with which to evaluate the market structures of perfect competition and monopoly.

Consumer surplus is the difference between the *maximum price* a consumer is prepared to pay and the *actual price* the consumer has to pay. In a competitive

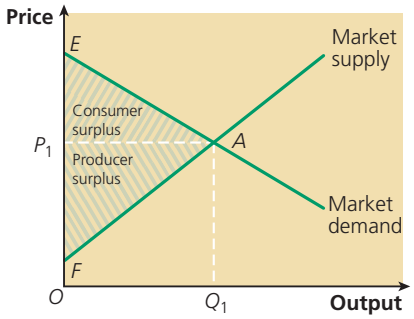


Figure 5.22 Consumer surplus and producer surplus

market illustrated in Figure 5.22, the total consumer surplus enjoyed by all the consumers in the market is measured by the triangular area P_1EA. Consumer welfare increases whenever consumer surplus increases — for example, when market prices fall. Conversely, however, higher prices reduce consumer surplus and welfare.

Producer surplus, which is a measure of producers' welfare, is the difference between the *minimum price* a firm is prepared to charge for a good and the *actual price* charged. In Figure 5.22, the producer surplus enjoyed by all the firms in the market is measured by the triangular area P_1FA.

It is important to understand consumer surplus and producer surplus in order to analyse how economic welfare may be affected by events that raise or lower the price of a good.

KEY TERMS

producer surplus a measure of the economic welfare enjoyed by firms or producers: the difference between the price a firm succeeds in charging and the minimum price it would be prepared to accept.

deadweight loss the loss of economic welfare when the maximum attainable level of total welfare fails to be achieved.

SYNOPTIC LINK

You will come across consumer surplus and producer surplus again in Chapter 10 when analysing the effect of an import duty or tariff on economic welfare.

Figure 5.23 shows how consumer surplus falls, and producer surplus increases, when the perfectly competitive industry is transformed into a monopoly — provided we assume that neither economies of scale nor dynamic efficiency improvements result from the transformation: that is, assuming that the marginal costs under monopoly and perfect competition are the same.

As Figure 5.23 shows, industry output in perfect competition is determined at point *A*, with output at Q_1 and price at P_1. However, monopoly equilibrium is determined at point *B* where *MR* = *MC*. (Note that the *marginal cost curve* in *monopoly* is the same curve as the *market supply curve* in *perfect competition*.) The diagram illustrates the standard case against monopoly, namely that compared to perfect competition, monopoly restricts output (to Q_2) and raises price (to P_2).

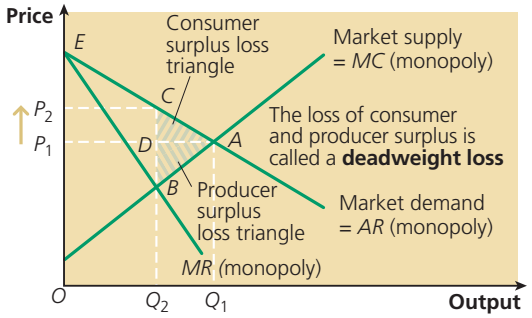


Figure 5.23 Consumer surplus and producer surplus in monopoly

We can use Figure 5.23 to illustrate how consumer surplus and producer surplus (and hence economic welfare) are affected by a monopoly replacing perfect competition. By raising the price from P_1 to P_2 the monopoly gains some of the consumer surplus that would have existed under perfect competition. The reduction in consumer surplus is equal to the rectangular area P_1P_2CD. This means that producer surplus (in the form of monopoly profit) increases at the expense of consumer surplus. Over and above this transfer, however, there is a net loss of economic welfare caused by the fact that the amount bought and sold falls to Q_2. The welfare loss or **deadweight loss** is shown by the two shaded triangular areas in Figure 5.23, which depict the loss of consumer surplus (the top triangle) and the loss of producer surplus (the bottom triangle). The deadweight loss is evidence of market failure in monopoly.

SYNOPTIC LINK

Imperfectly competitive market structures can lead to an inefficient or inequitable allocation of scarce economic resources. The resulting loss of welfare to society is a source of market failure. Chapter 8 discusses in more detail why the existence of monopoly may lead to market failure.

TEST YOURSELF 5.18

Explain why one of the four possible answers to the following question is correct and why the other three are wrong.

A decrease in the production costs of LCD television screens through the development of new technology is most likely to:

A raise price and decrease producer surplus

B maintain the existing price and increase consumer surplus

C lower price and increase consumer surplus

D maintain the existing price and decrease producer surplus

Consumer surplus and price discrimination

Firms undertaking price discrimination usually benefit from the practice, but their customers generally suffer. What is an advantage to firms or producers is a disadvantage for consumers. Figure 5.24 shows how consumer surplus falls when different prices are charged to different customers. Because consumer surplus is a measure of consumer welfare, consumers end up being worse off as a result of price discrimination. By contrast, the firms undertaking the price discrimination benefit. Their profits and producer surplus increase as a result of the transfer of consumer surplus away from their customers.

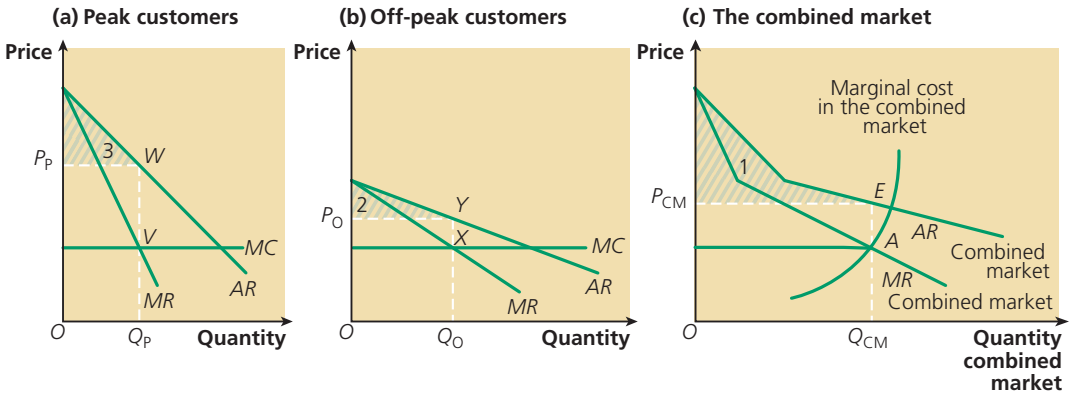


Figure 5.24 Price discrimination and the transfer of consumer surplus

As Figure 5.24 illustrates, price discrimination allows firms to increase profit by taking consumer surplus away from consumers and converting it into additional abnormal profit. Figure 5.24(a) and (b) are the same as Figure 5.19(a) and (b) on page 153. However, Figure 5.24(c) has been added to show the combined market with the peak and off-peak average revenue curves for rail travel added together. The peak and off-peak marginal revenue curves have also been added together. Note that for the combined market (but *not* the peak and off-peak sub-markets considered separately), the marginal cost curve slopes upwards, depicting the impact of the law of diminishing returns.

TEST YOURSELF 5.19

'Increasing consumer surplus is always good but increasing producer surplus is always bad.' Evaluate this statement.

In the absence of price discrimination, all rail travel consumers pay the same price, namely P_{CM} shown in Figure 5.24(c). Without price discrimination, consumer surplus is shown by the shaded area (labelled 1) above P_{CM} in panel (c). But with price discrimination, when peak customers are charged price P_P and off-peak customers P_O, consumer surplus falls to equal the shaded areas labelled 3 and 2 in Figure 5.24(a) and (b). The firm's profit has increased by

transferring consumer surplus from consumers to the producer. Producer welfare (or producer surplus) has increased at the expense of consumer welfare (or consumer surplus).

EXTENSION MATERIAL

Price discrimination: the limiting case

So far we have explained price discrimination where discriminating firms divide the market into a number of different market segments. This is also called third-degree price discrimination. Here we will look at first-degree price discrimination, which is also called 'perfect' price discrimination. It is the limiting case of price discrimination. First-degree price discrimination is illustrated in Figure 5.25. (There is also another type of price discrimination called second-degree price discrimination. This occurs when the price charged varies according to quantity demanded. With 'bulk buying', larger quantities are available at a lower price per unit — a 'bulk-buying' discounted price.)

Figure 5.25 illustrates a situation in which *all* the consumer surplus is transferred into producer surplus or producer welfare. Every customer is charged the maximum price each is prepared to pay. Figure 5.25 is basically the same as Figure 5.8 on page 136, which shows monopoly equilibrium. In the absence of price discrimination, the firm produces the level of output Q_1 where $MR = MC$ and all customers are charged the price P_1. Abnormal profit is shown by the rectangle C_1BAP_1, and consumer surplus by the shaded triangular area P_1AZ.

Now consider what happens when the firm charges each customer the maximum price each is prepared to pay. Customer Q_V depicted in the diagram is charged price P_V, customer Q_W is charged P_W, and so on. In this situation, there may be as many prices as there are customers. Because each customer is paying the maximum price he or she is prepared to pay, all the consumer surplus is transferred away from consumers to the firm, thereby boosting monopoly profit.

In summary, price discrimination leads to a loss of consumer surplus or consumer welfare. Firms exploit producer dominance and monopoly power, charging most consumers higher prices than would be charged in the absence of price discrimination. For these reasons, price discrimination is usually regarded as undesirable. However, Figure 5.25 also shows how *some* consumers at least, often the poorest consumers, can benefit from price discrimination. Each time the firm sells to one more customer, total sales revenue rises by the extra units sold multiplied by the price the customer pays. Because different customers are charged different prices, charging a high (or low) price to one customer does not affect the prices charged to other customers. In the absence of price discrimination, the firm's *AR* curve continues to be the demand curve the firm faces, with the firm's *MR* curve located below the demand (or *AR*) curve. But when each customer is charged the maximum price they are prepared to pay, the demand curve is now the same as the firm's *MR* curve. The profit-maximising level of output, where $MR = MC$, shifts to Q_2, located at point *Y* in Figure 5.25. Customers who would refuse to buy the good at price P_1 buy the extra output because the prices they are charged are lower than P_1. As a result, most consumers end up paying a price which is higher than P_1 (the profit-maximising price in the absence of price discrimination), but some consumers pay a lower price. The lowest of all the prices charged is P_2, which is the price charged to the marginal, and perhaps the poorest, customer.

Consider also a situation in which a firm cannot make enough profit to stay in business unless some consumer surplus is taken from consumers and transferred to the producer. Market provision of healthcare by a doctor in an isolated village or very small town is an example. When charging the same price to all her patients, a doctor cannot earn enough income to continue to provide the service. Without an increase in income, the doctor will move to a larger city and local medical care will no longer be available in the village. But with price discrimination, the rich pay a higher price than the poor. Everybody gets some benefit, and a needed service is provided.

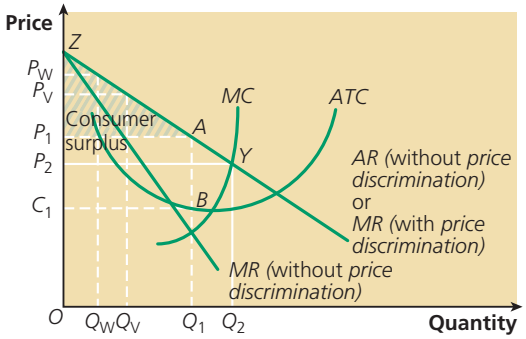


Figure 5.25 Price discrimination: the limiting case, when each customer is charged the maximum price he or she is prepared to pay

SECTION 5.12 SUMMARY

- The welfare concepts of consumer surplus and producer surplus can be used in the analysis of price discrimination and different market structures.
- Consumer surplus is the difference between the maximum price a consumer is prepared to pay and the actual price the consumer has to pay.
- Producer surplus, which is a measure of producers' welfare, is the difference between the minimum price a firm is prepared to charge for a good and the actual price charged.
- When perfect competition is replaced by monopoly, some of the consumer surplus that existed under perfect competition is transferred to the monopoly.
- This means that producer surplus (in the form of monopoly profit) increases at the expense of consumer surplus. Over and above this transfer, however, there is a net loss of economic welfare which is called a deadweight loss.
- The deadweight loss of welfare is evidence of market failure in monopoly.
- Price discrimination also transfers consumer surplus away from consumers and towards firms undertaking the price discrimination.
- However, price discrimination can enable firms to supply goods and services to poorer customers who otherwise would not be profitable to serve.

Questions

1 Explain why profit is maximised when $MR = MC$ both in perfect competition and in monopoly.

2 Explain why firms in real-world markets use methods of competition other than price competition.

3 Do you agree that perfect competition, if it existed, would always be preferable to monopoly? Justify your answer.

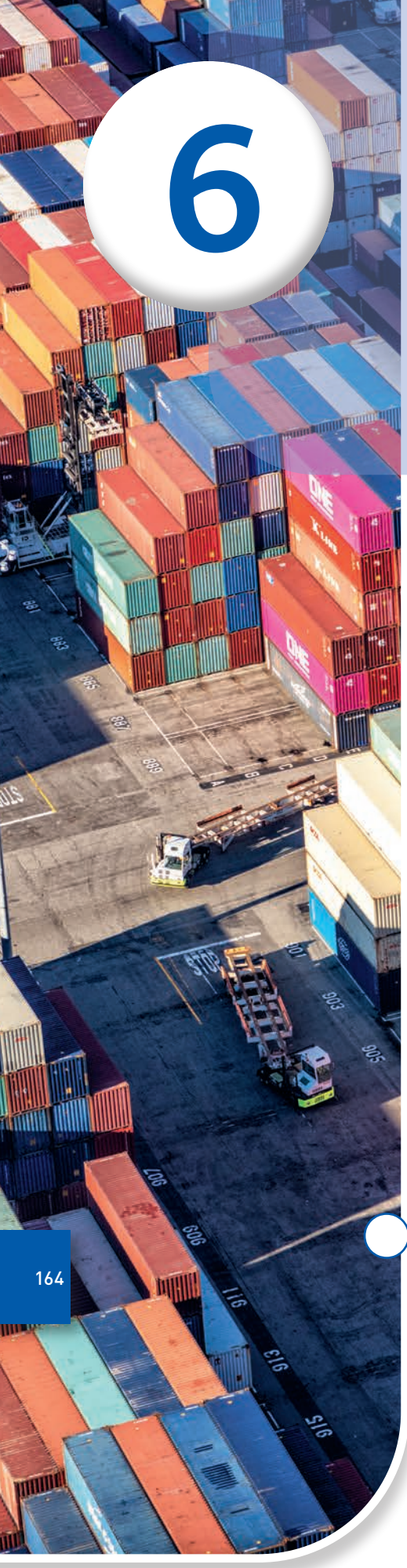
4 Distinguish between monopoly and monopoly power.

5 What is a cartel and why are cartels usually regarded as bad for the economy?

6 How can the process of creative destruction be related to changing market structures?

7 What is monopoly deadweight loss? Illustrate the concept on a diagram.

8 Explain how (a) firms and (b) consumers can benefit from price discrimination.



6 The labour market

This chapter reintroduces the analysis which occupied Chapter 3, namely the role of the price mechanism in allocating scarce resources between competing uses. However, unlike Chapter 4, which looked solely at markets for goods, or markets for outputs, this chapter focuses on the price mechanism and the labour market, which is a market for a factor of production or input into the production process. An important part of this chapter is the analysis of perfect competition and imperfect competition in the labour market. The imperfectly competitive labour market that we will examine is an example of monopsony. The chapter explains the similarities between, but also the differences that separate, perfect competition in the economy's goods market, studied in Chapter 5, and in its labour market. The chapter concludes by examining how trade unions, the national minimum wage and wage discrimination affect labour markets.

LEARNING OBJECTIVES

These are to understand:

- the similarities between labour markets and goods markets, but also the differences between these two sets of markets
- the demand for labour as a derived demand
- the demand for labour in a perfectly competitive labour market
- the supply of labour in a perfectly competitive labour market
- how wage rate and the level of employment are determined in a perfectly competitive labour market
- the similarities of and differences between a monopsony labour market and a monopoly goods market
- how wage rate and the level of employment are determined in a monopsony labour market
- the impact of trade unions on wage rates and employment
- the national minimum wage
- the effects of wage discrimination and other forms of discrimination in labour markets.

6.1 The demand for labour

The labour market and the goods market

Much of the theory we explain in this chapter, including the theory of the demand for labour, is really just the price theory you have already studied in the goods market of an economy, operating in the labour market. A labour market is an example of a factor market — that is, a market in which the services of a factor of production are bought and sold. Markets for land, capital goods and entrepreneurial skill are the other factor markets.

As Figure 6.1 shows, households and firms function simultaneously in both sets of markets, but their roles are reversed. Whereas firms are the source of supply in a goods market, in a factor market firms exercise demand for factor

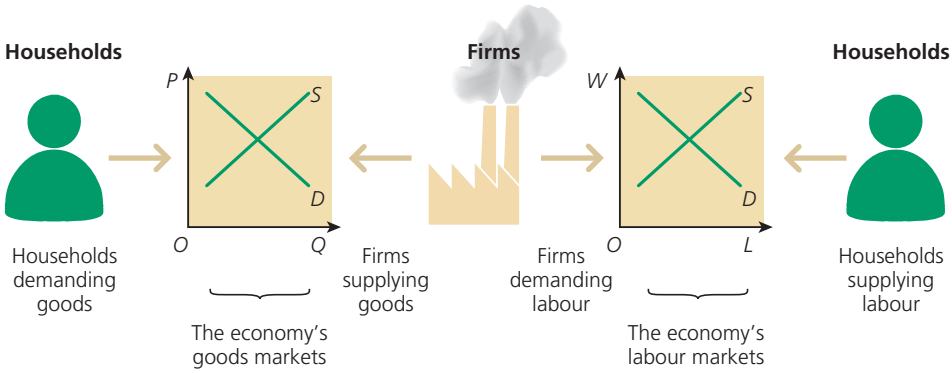


Figure 6.1 The goods market and the labour market, which is one of the economy's factor markets

services supplied by households. The incomes received by households from the sale and supply of factor services contribute in large measure to households' ability to demand the output supplied by firms in the goods market. To exercise demand, which requires an ability to pay as well as a willingness to pay, households need an income, and for most people this requires the sale of their labour services in a labour market.

The relationship between households and firms in the two markets is essentially circular. In goods markets, finished goods and services flow from firms to households, who spend their incomes on the goods. In labour markets, members of households earn the incomes they spend on goods by selling labour to their employers.

SYNOPTIC LINK
The relationship between households and firms in the economy's labour markets is central to the circular flow of income, which is explained in Chapter 10.

The demand for labour as a derived demand

A firm demands labour only if profits can be increased by employing more workers. But this assumes that households in goods markets demand the goods and services that workers are employed to produce. This means that a firm's demand for labour is a **derived demand** — it is derived from the demand for goods. Assuming a profit-maximising objective on the part of firms, there can be no demand for labour in the long run unless the firms employing labour sell the outputs produced for at least normal profit in the goods market.

KEY TERM
derived demand demand for a good or factor of production, wanted not for its own sake, but as a consequence of the demand for something else.

TEST YOURSELF 6.1
Name two goods or services other than labour for which demand is a derived demand.

STUDY TIP
Make sure you understand the difference between derived demand and composite demand. The concept of derived demand is crucial to understanding the demand for factors of production such as labour. If in doubt, see Chapter 3.

Marginal productivity theory and the demand for labour

The marginal physical product of labour

KEY TERM
marginal physical product of labour the addition to a firm's total output brought about by employing one more worker.

To understand a firm's demand curve for labour, we begin by explaining the **marginal physical product of labour** (MPP_L). Rather confusingly, this term is just another name for the marginal returns (or marginal product) of labour, which we explained in Chapter 4. MPP_L measures the amount by which a firm's total output rises in the short run (holding capital fixed), as a result of employing one more worker. In Chapter 4, we explained how the law of diminishing returns or diminishing marginal productivity operates as a firm employs more labour when capital is held fixed.

The law of diminishing returns is illustrated in Figure 6.2, which shows the marginal product of labour falling as additional workers are hired by the firm.

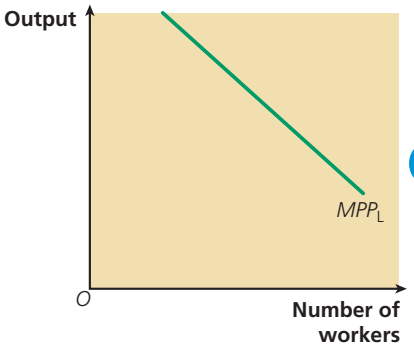


Figure 6.2 The marginal physical product of labour

SYNOPTIC LINK
Figure 6.2 assumes that labour suffers from diminishing marginal physical productivity whenever more labour is added to fixed capital. In other words, the possibility that at a low level of demand for labour, there could be an increasing marginal physical productivity of labour is ignored. We are assuming that the marginal physical product of labour falls whenever an additional worker is employed. Compare Figure 6.2 to Figure 4.4 in Chapter 4, which is not based on this assumption.

QUANTITATIVE SKILLS 6.1

Worked example: calculating the marginal physical product of labour

Using the total physical product data in Table 6.1, calculate the marginal physical product of labour of workers employed assembling wheelbarrows in a particular week.

Table 6.1 Total physical product of labour

Size of labour force	Total physical product of labour (wheelbarrows per week)
0	0
1	5
2	12
3	21
4	31
5	40
6	46
7	50
8	51
9	51
10	49

The marginal physical product of labour measures the change in the total physical product of labour when an additional worker is added to the labour force. For the first worker employed, the answer is calculated by subtracting the total physical output when no workers are employed from the total physical product of labour when one worker is employed. This is 5 – 0, which is 5. For the second worker, the answer is 12 – 5, which is 7. Likewise, for the 3rd, 4th, 5th, 6th, 7th, 8th, 9th and 10th workers, their marginal physical productivities are respectively: 9, 10, 9, 6, 4, 1, 0 and –2.

(If you were to plot this information on a graph, the slope of the marginal physical productivity curve would resemble the marginal returns curve drawn in the lower panel of Figure 4.4 in Chapter 4. This is because the total physical product schedule in Table 6.1 assumes that the wheelbarrow manufacturing business initially benefits from the increasing marginal productivity of labour before diminishing marginal returns set in.)

KEY TERM
marginal revenue product of labour the money value of the addition to a firm's total output brought about by employing one more worker.

The marginal revenue product of labour

The falling MPP_L curve in Figure 6.2, which is also drawn in Figure 6.3(a), shows the *physical* output produced by an extra worker — measured, for example, in cars or loaves of bread, or whatever goods or services the firm produces. To convert the marginal *physical* product of labour into the **marginal revenue product of labour** (MRP_L), the *MPP* of labour must be multiplied by marginal revenue (*MR*), which is the addition to the firm's total sales revenue resulting from the sale of each of the extra products in the goods market. Assuming a perfectly competitive goods market (as well as a perfectly competitive labour market), *MR* is identical to the good's price, and is shown by the horizontal *MR* curve in Figure 6.3(b).

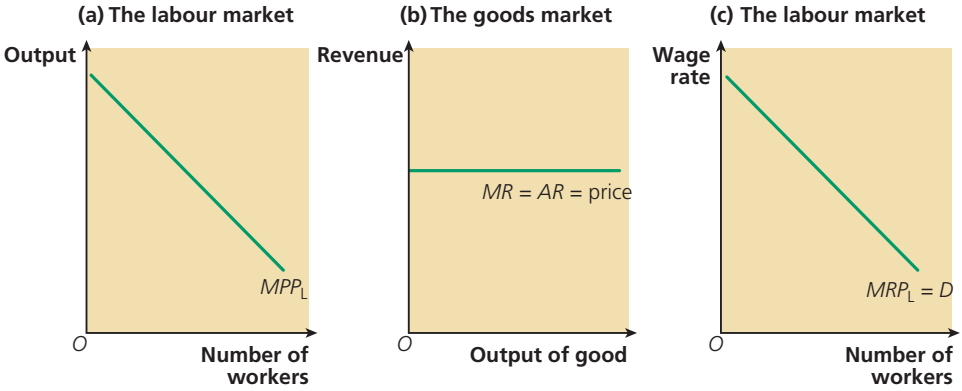


Figure 6.3 Deriving the MRP_L curve from the MPP_L curve

The marginal revenue product curve of labour, which is shown in Figure 6.3(c), can be explained with the use of the following equation:

marginal physical product × marginal revenue = marginal revenue product of labour

or:

$$MPP_L \times MR = MRP_L$$

Under perfect competition, since $MR = AR$ in the goods market where a firm sells its output, then in the labour market:

$$MRP_L = MPP_L \times \text{price}$$

SYNOPTIC LINK
Refer back to Chapter 4, page 104, for a full explanation of marginal revenue in a perfectly competitive goods market.

TEST YOURSELF 6.2
Distinguish between the marginal physical productivity and the marginal revenue productivity of labour.

EXTENSION MATERIAL

Deriving the *MRP* curve of labour when the goods market is imperfectly competitive

Assuming perfect competition in the goods market in which the firm sells its output, the slope of the *MRP* curve in Figure 6.3(c) is explained solely by the diminishing marginal physical product of labour: that is, by the law of diminishing returns.

However, if output is sold in an imperfectly competitive goods market, the marginal revenue product of labour declines faster than in a perfectly competitive goods market. This is because in imperfectly competitive goods markets, the marginal revenue earned from selling an extra worker's output also falls as output increases. The firm faces a downward-sloping demand curve for its products and, therefore, it has to reduce the price of the product in order to sell the extra output the worker produces. In this situation, there are *two* reasons for the MRP_L curve to fall as employment increases.

SYNOPTIC LINK

Refer back to Chapter 4 for explanations of the slope of the *MR* curve in perfectly competitive and monopoly goods markets.

TEST YOURSELF 6.3

Using the answers to Quantitative Skills 6.1, calculate the marginal revenue productivity of each of the ten workers, given the assumption that, however many wheelbarrows are produced, their price is always £60.

What a demand curve for labour shows

The demand curve for labour shows the relationship between the wage rate and number of workers employed. The lower the wage rate, the more labour a single employer, or all the employers, within a labour market will be willing to take on or hire.

Whenever there is more than one employer in the labour market, the market demand curve for labour is in a different position from a single firm's demand curve for labour. Figure 6.4(a) shows a *firm's* demand curve for labour in a perfectly competitive labour market. In such a market, the *market* demand curve for labour, which is shown in Figure 6.4(b), would simply be the addition of the demand curves for labour of each of the firms that make up the market. Since each firm's demand for labour would be shown by the MRP_L curve facing the firm, the market demand curve for labour would be the horizontal sum of all the MRP_L curves.

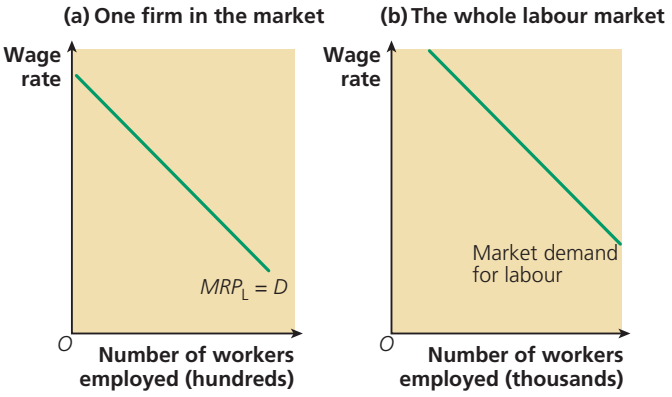


Figure 6.4 The *market* demand curve for labour in panel (b) is the sum of each *firm's* demand curve for labour in panel (a)

Shifts of the market demand curve for labour in different labour markets

TEST YOURSELF 6.4
Suggest two other causes of a rightward shift in the demand for labour.

The market demand curve for labour can shift to the right or left for a number of reasons. (A rightward shift is called an *increase* in demand for labour and a leftward shift is a *decrease* in demand for labour.) The possible reasons for a shift in the demand curve for labour are described below.

A change in labour productivity

As explained earlier, a firm's demand curve for labour is the *MRP* curve of labour facing the firm. It follows from this that if labour productivity increases, more labour is demanded at each wage rate and the firm's demand curve for labour shifts to the right. The same is the case when marginal revenue increases. (Remember, the demand for labour is a *derived demand*. If demand increases for the good that labour produces, raising the price of the product, the demand for labour curve shifts to the right.) And if a firm's demand curve for labour shifts to the right, the market demand curve for labour, which is the sum of each firm's demand curve for labour, must also shift to the right.

STUDY TIP
Don't confuse a *movement* or *adjustment along* a demand curve for labour with a *shift* of the demand curve for labour. In the former case, a fall in the wage rate relative to the price of capital causes more labour to be demanded because workers become relatively cheaper than capital to employ. Factor substitution occurs with labour replacing capital in the production process. In the latter case, an increase in the derived demand for labour leads to more labour being demanded at all wage rates, causing a rightward shift in the demand curve for labour.

A change in technology

Changes in technology can cause the demand curve for labour to shift, in some cases to the right, but in other cases to the left.

- If technical progress makes labour more productive relative to other factors of production, firms are likely to substitute labour for other factors of production, in which case the demand curve for labour shifts to the right. Technical progress is also likely to improve productive and dynamic efficiency, which reduce the cost of making the product; if this leads to lower prices and higher demand for the product, it is also likely to increase the demand for labour.
- However, technical progress can also have the opposite effect by causing firms to substitute capital for labour. For example, a switch to automated methods of production, which increases the productivity of capital relative to labour, may reduce the demand for labour, thereby causing the demand curve for labour to shift to the left. Technical progress can also increase the demand for certain types of labour at the expense of workers with other skills, who may lose their jobs. The technological unemployment which results is a part of the process of creative destruction which we described in Chapter 4, on page 116.

The elasticity of demand for labour

If the wage rate increases, but nothing else that might affect the demand for labour changes, by how much will employment fall? The answer is affected by firms' wage **elasticity of demand for labour**. The demand for a particular type of labour is likely to be relatively inelastic:

- when the relevant wage cost forms only a small part of total production costs (this has been called 'the importance of being unimportant')
- when the demand for the good or service being produced by the labour is inelastic
- when it is difficult to substitute other factors of production, or other types of labour, for the labour currently employed
- in the short run, rather than the long run, since it often takes time for employers to adjust the method of production.

KEY TERM
elasticity of demand for labour proportionate change in demand for labour following a change in the wage rate. The elasticity can be calculated by using the following formula:

percentage change in quantity of labour demanded / percentage change in the wage rate

QUANTITATIVE SKILLS 6.2

Worked example: calculating the effect of a wage increase on employment

You are a labour market analyst providing advice to a trade union, whose members are unskilled workers currently being paid £10 an hour for their work. The union, which is pushing for a wage increase of 50p per hour, wants to know what will happen to the employment of its 20,000 members who currently have jobs. Your research tells you that the employer's wage elasticity of demand for unskilled workers is –10.

Assuming that everything else in the economy remains unchanged, calculate the effect of an increase in the wage rate by 50p per hour on the employment of the union members.

Given that the wage rate initially is £10 an hour, the proposed wage rate increase of 50p an hour is a 5% increase. The formula for the employer's wage elasticity of demand for labour is:

$$\text{wage elasticity of demand for labour} = \frac{\text{percentage change in quantity of labour demanded}}{\text{percentage change in the wage rate}}$$

% change in quantity of labour demanded = 5% × –10 = –50%

Given that 20,000 unskilled union members currently have jobs, the data indicate that half of them (10,000 union members) might lose their jobs.

STUDY TIP

Remember to avoid confusing the *slope* of a demand or supply curve with its *elasticity*. As noted in Chapter 3, the elasticity of demand will vary along a demand curve.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.1

Labour shortages in the fruit-picking market

In the summer of 2018, three out of five soft-fruit growers were struggling to recruit the 30,000 seasonal staff needed. And with a shortage of seasonal workers to pick the fruits, a trend exacerbated by the fall in inward migration of fruit pickers from central Europe caused by Brexit, there were growing fears that many soft fruits would be left to rot on their stems. Growers said that they expected the position to be worse in future years. The labour market was definitely tightening as the demand for workers exceeded their supply.

Soft-fruit production in the UK grew by 131% between 1998 and 2018, largely as a result of an increase in home-grown strawberries, which are now protected from bad weather by plastic sheeting. The industry was worth more than £1.2 billion in 2018. The staff shortage was not restricted to fruits, with the entire horticultural sector badly hit by Brexit and increased demand for labour in eastern Europe. As unemployment has fallen in Bulgaria and Romania, the number of people applying to work in England's fields has dropped, with the UK's departure from the European Union now amplifying the trend.

It has led to uncertainty over future immigration rules, while fears of encountering a growing hatred of immigrants and a weakened British currency — resulting in less money to send home — have also deterred arrivals. One fruit picker said that her friends



Many soft-fruit growers have traditionally relied on eastern European labour for fruit picking

don't go to the UK any more, but to Italy and Germany. The labour squeeze is occurring despite workers being mostly paid at the national living wage of £7.83 an hour (at the time) and some given airline tickets, housing support and productivity bonuses.

The prospect of British fruit pickers replacing foreign workers is remote. 'We had two British applicants this year, in six months,' said the boss of one recruitment agency, adding that there was absolutely no appetite for the jobs among British workers.

Farmers are calling on the government to open the immigration system to more seasonal workers from countries outside the EU, as Germany, Spain, Portugal and Ireland have already done. The UK, whose citizens voted to leave the EU to put a brake on European immigration, is ironically perhaps the only member country to recruit only within the EU.

Without government intervention, the impact of the labour shortage on the soft-fruit industry could be devastating. A third of British producers had already decided to scale back investment, while others were looking at expanding in different countries.

Follow-up questions

1 Why is the demand for fruit pickers highly seasonal?
2 What is meant by the statement 'The labour market was definitely tightening'?
3 Explain why the national living wage, paid at £7.83 an hour at the time, apparently had no effect on the labour shortage.
4 Evaluate the effect of the 2016 Brexit vote on the agricultural labour market in the UK.

SECTION 6.1 SUMMARY

- A labour market is an example of a factor market — that is, a market in which the services of a factor of production are bought and sold.
- In labour markets, members of households earn the incomes they spend on goods by selling labour to their employers.
- The demand for labour is a derived demand.
- Assuming perfect competition, the marginal revenue product curve is the firm's demand curve for labour.
- The elasticity of demand for labour is the proportionate change in the demand for labour following a change in the wage rate.
- The market demand curve for labour is the horizontal sum of the MRP_L curves facing all the firms in the labour market.
- The market demand curve for labour can shift to the right or left for a number of reasons.
- The reasons include a change in labour productivity and a change in technology.

6.2 Influences on the supply of labour to different labour markets

The market supply of labour

Just as the market demand curve for labour shows how much labour all the firms in the labour market plan to demand at different wage rates, so the market supply curve of labour shows how much labour all the workers in the labour market plan to supply at different wage rates. Figure 6.5 shows the market supply curve of labour.

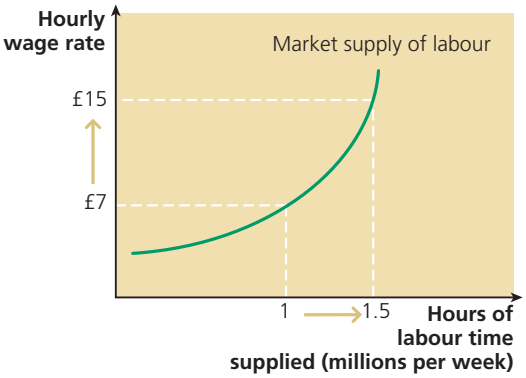


Figure 6.5 A market supply curve of labour in a particular week

The graph shows how many hours of labour workers plan to supply — for example, at different hourly wage rates — in a particular occupation in a particular week. The diagram depicts a situation in which workers respond to an increase in the hourly wage rate from £7 to £15 by increasing the total number of hours they are prepared to work from 1 million hours to 1.5 million hours per week.

Monetary and non-monetary considerations

SYNOPTIC LINK
For most people, both money and leisure time are affected by the law of diminishing marginal utility, which is explained in Chapter 2 on page 27.

The money wage rate on offer is the first of two factors which influence the amount of labour that workers are willing to supply. The second factor is the utility or economic welfare derived from other aspects of working, sometimes called the non-monetary benefits (these include what is popularly known as job satisfaction or, if negative, job dissatisfaction).

Different types of work yield different amounts of job satisfaction and dissatisfaction. When a worker enjoys the job, the net advantage of work is greater than the welfare yielded by the wage. In this situation, the worker is willing to work for a money wage lower than the wage that would be acceptable if there were no satisfaction from the work itself. But work such as routine assembly-line work in factories and heavy manual labour is unpleasant, yielding job dissatisfaction for some workers. The supply of labour for this type of employment reflects the fact that the hourly wage rate must be high enough to compensate for the unpleasantness (or sometimes the danger) of the job.

Figure 6.6 shows a worker who is free to decide how many hours to work each day or each week, choosing between supplying more labour or enjoying more leisure time. As more labour time is supplied at a particular hourly wage rate, the extra income yields less and less extra satisfaction for the worker. But because there are only 24 hours in the day, the decision to supply more labour simultaneously means the decision to enjoy less leisure time. In this situation, each extra hour of leisure sacrificed is accompanied by an increasing loss of economic welfare. At the margin, to maximise personal welfare, a worker must supply labour up to the point at which:

utility or welfare from the last unit of money earned = utility or welfare from the last unit of leisure time sacrificed

In this situation, the marginal private benefit received by a worker from supplying labour equals the marginal private cost incurred from giving up

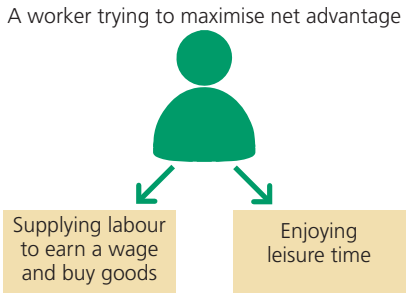


Figure 6.6 The choice between supplying labour and enjoying leisure time

leisure time. Provided personal preferences remain stable, there is no incentive for the worker to supply more labour at the *going* hourly wage rate.

However, a *higher* hourly wage will provide an incentive to work more hours. With a higher wage rate, at the margin, the welfare derived from the wage becomes greater than the welfare derived from the last unit of leisure time enjoyed. To maximise personal welfare at the higher wage rate, the worker would be expected to supply more labour and enjoy less leisure time. The result is the upward-sloping labour supply curve shown earlier in Figure 6.5. An increase in the wage rate from £7 an hour to £15 an hour increases the number of hours worked in the labour market from 1 million hours to 1.5 million hours.

TEST YOURSELF 6.5
Explain what is meant by economic welfare.

Individual workers' job choice depends on the utility they expect to derive from the money wage rate and job satisfaction or dissatisfaction in different occupations. However, given that the non-monetary advantages usually change slowly, a rise in the wage in an occupation should lead to an increase in the supply of labour.

SYNOPTIC LINK
At this point in the course, it is a good idea to remind yourself of the meaning of opportunity cost. As Chapter 1 explains, the opportunity cost of any choice, decision or course of action is measured in terms of the alternatives that have to be given up. There is an opportunity cost whenever a person decides to work. The opportunity cost of supplying one more hour of labour time (in order to earn money) is the hour of leisure time sacrificed. Because of the time constraint resulting from the fact that there are only 24 hours in a day, a decision to supply one more hour of labour time means that the worker chooses one hour less of leisure time. Labour time and leisure time are substitutes for each other, and working longer hours eats into leisure time.

Shifts of the market supply curve of labour

Changes in the non-monetary factors affecting the supply of labour shift both an individual worker's supply curve of labour and the market supply curve of labour to a new position. These factors include the benefits of job security, promotion prospects, good working conditions, holiday entitlement and other psychological benefits of work. Improvements in these benefits will shift the supply curve of labour to the right. A deterioration in these benefits shifts the supply curve of labour to the left. Generalising, if people decide they value leisure more highly, they will work fewer hours at each wage rate, and the supply curve for labour will shift to the left. If they decide they want more goods and services rather than leisure time, the supply curve is likely to shift to the right.

Other factors that can shift the supply curve of labour include:

- **Changes in income.** It is reasonable to assume that leisure time is a normal good for most people, so a rise in income increases the demand for leisure time, which in turn causes the supply curve of labour to shift to the left. However, for a few people, leisure time might be an inferior good. For them, higher income reduces their demand for leisure time and they end up supplying more labour.
- **Changes in population.** A rise in population, perhaps caused by immigration, increases the supply of labour; a reduction lowers it. A fall in the number of people of working age causes the labour supply curve to shift to the left, except perhaps when this is offset by those who have reached retirement age deciding to work longer in order to finance their eventual retirement.

TEST YOURSELF 6.6
Suggest two factors which might cause the market supply curve of labour to shift to the right.

- **Changes in expectations.** Likewise, if older people expect to live longer yet become less optimistic about their future pensions, this could increase the labour supply. A rise in the proportion of people staying on in further and higher education will tend to reduce the supply of labour.

SYNOPTIC LINK

At this point you might refer back to Chapter 3 and refresh your understanding of normal goods and inferior goods. This is because changes in the prices of such goods will have differing impacts upon the corresponding demand for labour.

The elasticity of supply of labour

KEY TERM

elasticity of supply of labour proportionate change in the supply of labour following a change in the wage rate. The elasticity can be calculated by using the following formula:

percentage change in quantity of labour supplied / percentage change in the wage rate

The factors which determine the wage **elasticity of supply of labour** include the following:

- The supply of unskilled labour is usually more elastic than the supply of a particular type of skilled labour. This is because the training period of unskilled labour is usually very short, and any innate abilities required are unlikely to be restricted to a small proportion of the total population.
- Factors which reduce the occupational and geographical mobility of labour tend to reduce the elasticity of labour supply. Occupational mobility of labour is determined by how easily workers can change professions, while geographical mobility relates to how easily workers can move to where the current jobs are located.
- The supply of labour is also likely to be more elastic in the long run than in the short run.
- The availability of a pool of unemployed labour increases the elasticity of supply of labour, while full employment has the opposite effect.

QUANTITATIVE SKILLS 6.3

Worked example: calculating the wage elasticity of supply

Fearing a shortage of labour, a firm that wants to expand, and which currently employs 1,000 workers, increases the wage it pays to all its workers from £10 per hour to £15 per hour. As a result, 500 workers apply for jobs at the firm.

What is the wage elasticity of supply of labour with respect to the stated wage rate increase?

The formula needed to calculate the correct answer is:

$$\text{wage elasticity of supply of labour} = \frac{\text{percentage change in quantity of labour supplied}}{\text{percentage change in the wage rate}}$$

$$= +50\% / +50\% = +1$$

Thus, between these two hourly wage rates the wage elasticity of supply of labour is unitary: that is, equal to 1.

TEST YOURSELF 6.7

Distinguish between the formulas for the elasticity of demand for labour and the elasticity of supply of labour.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.2

Who should be paid more, MPs or newsreaders?

Economics teachers often ask their students to ponder on whether workers employed in pleasant occupations, such as television celebrities, should be paid more than those in disagreeable occupations, such as road sweepers. If job satisfaction or dissatisfaction were to be the only factor determining wages, the road sweeper would be paid more. However, more often than not, other factors, related to supply and demand, productivity, learned skills and innate ability, tend to override the job satisfaction factor and explain why television celebrities are paid more than road sweepers.

A few years ago, a government minister asked: 'should MPs be paid more than television newsreaders?' In a follow-up poll, 45% of the general public said 'yes' and 55% voted 'no'. Former Conservative MP Michael Portillo, now a TV presenter of programmes about railway journeys, argued that:

> low pay will discourage good people from becoming MPs, especially if we are now going to denigrate their outside earnings too, even though that's how TV presenters boost their salaries. The present exchange rate is approximately one Zoe Ball equals 14 MPs. It is a topsy-turvy world where we pay top journalists more to comment occasionally on what the prime minister does than we pay him to do his job 24/7.

Replying 'no', Ken Livingstone, an ex-mayor of London and former Labour MP, said:

> Far too many people are paid too much for what they do. In 1979, Britain's top 10% of earners were paid four times as much as the bottom 10%. By 2018 that figure had doubled to at least ten times as much. Is Britain more than twice as well run as it was 40 years ago? Is the output of television twice as good as in the 1970s? Are our politicians and bankers twice as honest? It's time for a dramatic reduction in top salaries across the board, not just those of TV presenters.

Follow-up questions

1. Do you believe that both MPs and TV newsreaders are paid too much? Use economic theory to justify your argument.
2. Name two other groups of workers whom you think are paid too much. Justify your answer.
3. Name two other groups of workers whom you think are paid too little. Justify your answer.
4. Should relative wage and salary rates be determined solely by market forces? Again, justify your answer.

SECTION 6.2 SUMMARY

- The market supply curve of labour slopes upwards, showing that workers plan to supply more labour as the wage rate rises.
- The supply of labour is affected by non-monetary considerations or rewards as well as by monetary considerations or rewards.
- Non-monetary rewards include job satisfaction.
- Changes in non-monetary rewards shift the supply curve of labour to a new position.
- Likewise, changes in income, population size and expectations shift the position of the supply curve of labour.
- The opportunity cost of working is the leisure time forgone.
- The elasticity of supply of labour is the proportionate change in the supply of labour following a change in the wage rate.

6.3 Perfectly competitive labour markets

In this section, we bring the demand for, and supply of, labour together and explain the determination of relative wage rates and levels of employment in perfectly competitive labour markets.

A perfectly competitive *labour* market, if it were to exist, would have to meet all of the following conditions at the same time. As would be the case in a perfectly competitive *goods* market, a perfectly competitive labour market would have to contain a large number of buyers and sellers, each unable to influence the ruling market price (in this case, the ruling market wage), and operating in conditions of perfect market information. Employers and workers would be free to enter the labour market in the long run, but an individual employer or firm could not influence the ruling market wage through its independent action.

Just as in a perfectly competitive goods market, it is impossible for all these requirements to be met simultaneously. It follows that perfectly competitive labour markets do not exist in the real world. Some labour markets, such as the market for fruit pickers in a region or district where there are a very large number of orchards, *approximate* to perfect competition, but nevertheless they are not perfectly competitive.

SYNOPTIC LINK

Refer back to Figure 5.1 in Chapter 5, which lists the six conditions which would have to be met for perfect competition to exist.

How the ruling market wage affects perfectly competitive firms

KEY TERMS

average cost of labour total wage costs divided by the number of workers employed.

marginal cost of labour the addition to a firm's total cost of production resulting from employing one more worker.

In Chapter 5, we explained how a firm in a perfectly competitive goods market would be able to sell as much as it wanted at the ruling market price, meaning that the firm faces a perfectly elastic demand curve which is also the firm's average and marginal revenue curves. Each firm is a passive price-taker at the ruling price determined in the market as a whole, choosing the quantity to sell, but not the price.

A very similar situation would exist for a firm employing workers in a perfectly competitive labour market, except that now the firm could *buy* as much labour as it wanted to employ at the ruling market wage. To state this another way, each employer would face a perfectly elastic *supply* of labour curve in a perfectly competitive labour market. Figure 6.7 illustrates why.

In a perfectly competitive labour market, each employer would have to passively accept the ruling market wage. The ruling wage, determined in Figure 6.7(b), is also the perfectly elastic supply curve of labour facing each of the firms in the labour market. This means it is also the **average cost of labour** (AC_L) curve and the **marginal cost of labour** (MC_L) curve facing each firm in the labour market. This is shown in Figure 6.7(a).

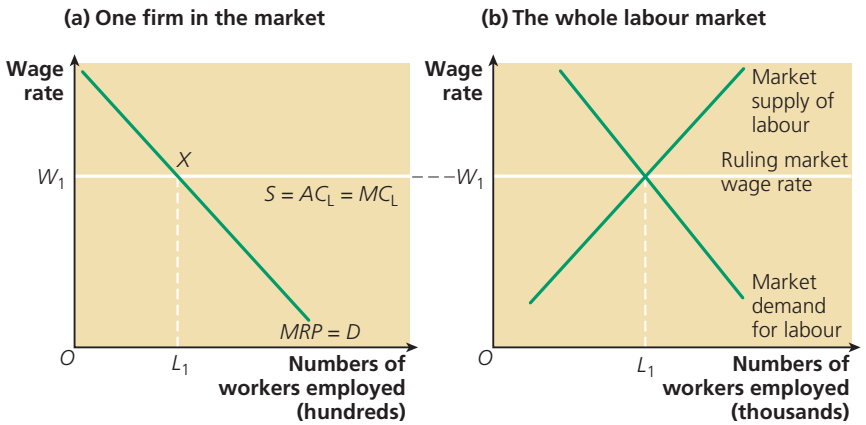


Figure 6.7 A perfectly competitive labour market

At the ruling market wage, each firm would be a passive price-taker, able to hire as many workers as it wished to hire at the ruling market wage, but unable to influence the ruling wage by its own actions.

Equating MRP_L with the wage rate

To maximise profit when selling the output produced by labour, each firm would have to demand labour up to the point at which:

the addition to sales revenue resulting from employment of an extra worker	=	the addition to production costs resulting from the employment of an extra worker

or:

$$MRP_L = MC_L$$

The marginal revenue product of labour (MRP_L) is the marginal benefit accruing to the employer when hiring an extra worker. Likewise, the marginal cost of labour (MC_L) is the marginal private cost incurred by each firm. Since, in a perfectly competitive labour market, the marginal cost of labour would always equal the wage paid to the workers, the perfectly competitive firm's level of employment or demand for labour at each wage rate would be where:

$$MRP_L = W$$

Point X in Figure 6.7(a) shows the number of workers that a firm would be willing to employ at a ruling wage rate W_1. Consider what would happen if the firm were to employ a labour force larger than L_1. Additional workers would add more to total costs of production than to total revenue, and so total profit would fall. Conversely, with a workforce below L_1, the *MRP* of the last worker is greater than the wage, and the total profit would increase if more workers were employed.

Summarising:

- If $MRP > W$, more workers should be hired.
- If $MRP < W$, fewer workers should be employed.
- If $MRP = W$, the firm is employing the number of workers consistent with profit maximisation.

As the *MRP* curve shows each firm's demand for labour at each possible wage rate, including wage rate W_1, the *MRP* curve facing each firm would in fact be the firm's demand curve for labour.

TEST YOURSELF 6.8
What is meant by profit maximisation?

The role of market forces in determining relative wage rates

In a perfectly competitive labour market, each employer would be just one among many in the market, able to hire however many workers it wished to employ, provided only that the ruling market wage was offered to all employees taken on.

Look again at Figure 6.7(b). This shows the equilibrium market wage W_1 determined where the market demand curve for labour intersects the market supply curve of labour. Once market forces, operating in the whole labour market, have established the ruling market wage, each worker in the labour market is a passive wage-taker at the market-determined wage rate.

Refer back also to Application of Economics in the Real World 5.2 in Chapter 5 (page 130), which mentions the 'law of one price'. The law says that identical goods should sell for the same price in two separate markets, the more nearly perfect both markets are. Applied to the labour market, this means that, if labour markets were perfectly competitive and in equilibrium, workers with the same skills and performing similar tasks would be paid the same wage rates.

Of course, in the real world this does not happen. Real-world markets are imperfectly competitive to a greater or lesser extent. And even if most of the conditions of perfect competition were met, different skills and productivities, together with changes in demand for different groups of workers, would mean that differences in wage rates would persist.

SECTION 6.3 SUMMARY

- In this section we have brought together the demand for, and supply of, labour so as to explain the determination of relative wage rates and levels of employment in perfectly competitive labour markets.
- To understand a perfectly competitive labour market, you need to appreciate the similarities of such a market to a perfectly competitive market for a good such as wheat, but also the differences between goods and labour markets.
- The supply curve of labour facing an individual firm in a perfectly competitive labour market is the horizontal line depicting a perfect elasticity of supply, determined by the interaction of supply and demand in the market as a whole.
- For a firm in a perfectly competitive labour market, this horizontal line is also the average cost of labour (AC_L) and the marginal cost of labour (MC_L).
- To maximise profit when selling the output produced by labour, each firm would have to demand labour up to the point at which the addition to sales revenue resulting from employment of an extra worker equalled the addition to production costs resulting from the employment of an extra worker.

6.4 Imperfectly competitive labour markets

Just as is the case with markets for goods, all real-world labour markets are imperfectly competitive to a lesser or greater degree. At one extreme, some labour markets approximate to perfect competition, but at the opposite extreme are monopsonistic labour markets.

Monopsony means a single buyer, just as monopoly means a single seller. The term 'monopsony' was first used by the economist Joan Robinson in *The Economics of Imperfect Competition* (1933), to describe single, powerful buyers in product markets and/or labour markets. In a pure monopsonistic

TEST YOURSELF 6.9
Explain the difference between monopsony and monopoly.

KEY TERM
monopsony there is only one buyer in a market.

TEST YOURSELF 6.10
Why are many UK labour markets better described as monopsonistic rather than as pure monopsonies?

labour market, workers cannot choose between alternative employers, since there is only one firm or employer available to hire their services. However, pure monopsonistic labour markets are extremely rare. The armed services are perhaps the nearest example of a pure monopsony, though even here we can argue that soldiers can sell their labour services to foreign-based militias or to the police or security guard firms. Nevertheless, many large firms exercise a high degree of **monopsony power** in the labour market in which they hire their workers.

KEY TERM
monopsony power the market power exercised in a market by the buyer of a good or the services of a factor of production such as labour, even though the firm is not a pure monopsonist.

Monopsony power refers to a situation in which there are dominant employers but not pure monopsony. In the provision of healthcare, the National Health Service has considerable monopsony power, though the existence of private hospitals shows that the NHS is not a pure monopsony.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.3

Recession, recovery and the UK jobs market

When the UK economy entered recession in 2008 as a result of the financial crisis that started in 2007, many UK companies, including the construction plant manufacturers JCB, began to sack workers. Those workers who still had jobs feared they would soon be made redundant. To try to prevent further lay-offs, union members at JCB offered to work for lower wages. However, wage reductions did not work. In fact, JCB sacked even more workers as orders for bulldozers and other construction plant dried up.

However, by 2018, JCB was creating jobs again as the UK economy recovered from recession. In June 2018 the company announced an investment of more than £50 million in a new British plant, which it hopes will create hundreds of jobs. Many of the workers JCB hires are agency workers and lack the guarantee of continuous employment. However, JCB says that most of its full-time employees have entered through some period of temporary work and proved their worth. With an improving economy, the company has taken them on. And as a result of investing in its training academy, JCB has recently created 100 jobs for teenage apprentices. But JCB is creating far more jobs by investing in factories in Brazil and India. A company spokesman has said, 'Growth for JCB anywhere is good for JCB and a significant number of the machines are still made here in the UK and shipped around the world.'



As a manufacturer of capital goods, JCB is especially vulnerable to changes in demand for the machinery it produces

Follow-up questions

1 What is a capital good, and why are companies such as JCB especially vulnerable to changes in demand?
2 With the help of a demand curve for labour, explain why JCB laid off workers in 2008.
3 Explain why companies such as JCB, whose clients are building companies, are especially likely to lay off workers in a recession.
4 In what way does the information provided here illustrate the globalisation process?

The determination of relative wage rates and levels of employment in a monopsony labour market

Various factors, such as monopsony power, trade unions and imperfect information, contribute to imperfections in a labour market. We will examine the role of trade unions in section 6.5. In this section we look at monopsony and monopsony power.

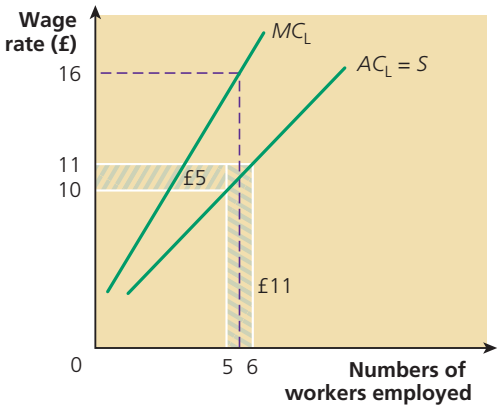


Figure 6.8 In a monopsony labour market, the MC_L curve lies *above* the AC_L or supply curve

Although in some ways a monopsonistic labour market resembles a monopolistic goods market, there are also significant differences. In much the same way that the market demand curve facing a monopoly supplier of a good is the monopolist's average revenue curve, so in a monopsonistic labour market, the market supply curve of labour is the firm's average cost of labour curve. The AC_L curve shows the different wage rates that the monopsony must pay to attract labour forces of different sizes. For example, Figure 6.8 shows a monopsonistic employer hiring five workers at an hourly wage rate or AC_L of £10. As the diagram shows, the hourly wage per week must rise to £11 to attract a sixth worker into the firm's labour force.

The supply or AC_L curve facing the monopsony shows the wage that has to be paid to *all* workers at each size of the labour force, to persuade the workers to supply their services. However, in a monopsony labour market, the AC_L curve is *not* the marginal cost of labour curve (MC_L). To attract an extra worker, the monopsony must raise the hourly wage rate and pay the higher wage to *all* the workers. In this situation, the marginal cost of labour incurred by employing an extra worker is the change in total amount by which the wage bill rises, not just the wage rate paid to the additional worker hired. The MC_L curve of labour illustrated in Figure 6.8 is positioned *above* the AC_L (or supply) curve.

TEST YOURSELF 6.11
What is monopsony power?

SYNOPTIC LINK
If you refer back to Chapter 4, pages 105–06, you will see that for a similar reason, in the goods market, a monopolist's *MR* curve is *below* its *AR* curve.

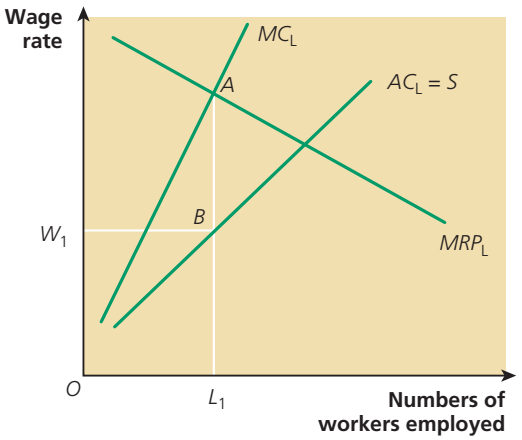


Figure 6.9 The determination of the wage rate and the level of employment in a monopsony labour market

In Figure 6.8, the MC_L incurred per hour by employing the sixth worker is £16, made up of the £11 wage rate paid to the sixth worker, plus the £1 extra that now has to be paid to each of the five workers already employed before the sixth worker joined the labour force.

Figure 6.9 shows how the equilibrium wage and the equilibrium level of employment are determined in a monopsonistic labour market. As in the case of a perfectly competitive employer, the monopsonist's level of employment is determined by the point where $MRP_L = MC_L$. This occurs at point *A* in Figure 6.9, with L_1 workers being hired.

However, the wage rate is determined at point *B* on the supply curve (and AC_L curve), which lies below point *A*. The monopsony

wage rate (W_1 in Figure 6.9) is therefore less than the marginal revenue product of labour. At point B, L_1 workers are willing to work for an hourly wage rate of W_1. Although the monopsony could pay a wage rate higher than W_1, it has no need to do so. Why pay more, when W_1 attracts all the workers the monopsony wants to hire? Indeed, if the monopsony were to pay a wage higher than W_1, it would inevitably incur unnecessary production costs and fail to maximise profits when selling its output in the goods market. Profit maximisation requires that L_1 workers are employed and a wage of W_1 is paid.

Reasons for wage differences in imperfectly competitive labour markets

In imperfectly competitive labour markets, wage differences in different labour markets are often substantial. Five reasons for this are:

- **Disequilibrium trading.** Economies are subject to constant change, such as the development of new goods and services and improved methods of production or technical progress. Patterns of demand also change. Because market conditions are always changing, labour markets — like other markets — are usually in disequilibrium rather than in equilibrium. Although market forces tend to equalise wages in competitive labour markets, at any point in time disparities exist, reflecting the disequilibrium conditions existent at the time.
- **Imperfect market information.** As the name implies, imperfectly competitive labour markets are characterised by imperfect market information. Workers sometimes lack accurate information on rates of pay, not only in other labour markets, but also within the industry in which they are selling their labour. Likewise, employers lack information about wage rates in other labour markets and also, on occasion, within their own industries. Imperfect market information also contributes to the immobility of labour, which is described below.
- **Occupational immobility of labour.** This occurs when workers are prevented, by either natural or artificial barriers, from moving between different types of job. Workers are obviously not homogeneous or uniform, so differences in natural ability may prevent or restrict movement between jobs. Artificial or 'man-made' barriers also prevent workers from moving between labour markets. These barriers include membership qualifications imposed by professional bodies such as accountancy associations, and trade union restrictive practices such as closed shops, which restrict employment to union members. Non-members may find it difficult or impossible to join the trade union, though such practices are now illegal in the UK.
- **Geographical immobility of labour.** This occurs when factors, such as ignorance of job opportunities, family and cultural ties, and the financial costs of moving or travel, prevent a worker from filling a job vacancy located at a distance from his or her present place of residence. Perhaps the most significant cause of geographical immobility in the UK in recent years has been the state of the housing market, which itself reflects imperfections in other factor markets. During house price booms, low-paid and unemployed workers in the northern half of the UK have found it difficult or impossible to move south to fill job vacancies in the more prosperous southeast of England. The prices of owner-occupied housing have soared and there has been very little housing available at affordable rents in either the private or

TEST YOURSELF 6.12
Explain how the wage rate would be determined in an imperfectly competitive labour market.

KEY TERMS
occupational immobility of labour when workers are unwilling or unable to move from one type of job to another, e.g. because different skills are needed.

geographical immobility of labour when workers are unwilling or unable to move from one area to another in search of work.

the public sector. At the same time, workers living in their own houses in the southeast may be reluctant to apply for jobs elsewhere in the country, for fear that they will never be able to afford to move back to southern England.

- **Discrimination.** Various forms of racial, religious, age and gender discrimination affect both the demand for and the supply of labour, as discussed later in the chapter.

SYNOPTIC LINK
Discrimination in a labour market is explained in more detail in section 6.7.

TEST YOURSELF 6.13
State two examples of occupational immobility of labour and two examples of geographical immobility of labour.

SECTION 6.4 SUMMARY

- Real-world markets are imperfectly competitive to a greater or lesser extent.
- A monopsony is the only buyer of labour in an imperfectly competitive labour market.
- In monopsony, the wage rate and the level of employment are likely to be lower than in a perfectly competitive labour market.
- Wage differences in different labour markets can be caused by disequilibrium trading, imperfect market information, the occupational immobility of labour, the geographical immobility of labour, and discrimination in the labour market.

6.5 The influence of trade unions in determining wages and levels of employment

KEY TERMS
trade union a group of workers who join together to maintain and improve their conditions of employment, including their pay.
collective bargaining a process by which wage rates and other conditions of work are negotiated and agreed upon by a union or unions with an employer or employers.

A **trade union** is an association of workers formed to protect and promote the interests of its members. A major function of a union is to bargain with employers to improve wages and other conditions of work. In the analysis which follows, we will regard a trade union as a monopoly supplier of labour, which is able to keep non-members out of the labour market and also to prevent its members from supplying labour at below the union wage rate. Of course, in real life a union may not necessarily have the objectives specified above, and even if it does, it may not be able to achieve them.

In labour markets where trade unions are strong and possess significant bargaining power, wage rates are usually determined through a process known as **collective bargaining**. This term refers to a situation in which unions bargain with employers, with the unions negotiating collectively on behalf of their members. Sometimes bargaining takes place with a single employer, while on other occasions it may be with a number of employers in the labour market. The market wage rate is determined through the bargaining process, with the employers then deciding how many workers to employ at this wage rate.

However, many UK employers are now reluctant to recognise and bargain with the unions to which their employees belong. A number of factors are responsible for the shift in the balance of power in the labour market away from trade unions and towards employers. Two of the factors are a series of Employment Acts which have restricted the legal rights of trade unions, and the impact of globalisation and international competition on British labour markets. British trade unions generally remain powerful only in industries which are protected from international competition, such as London Transport, the rail industry and other areas of public-sector employment.

The decline of collective bargaining in the UK means that more and more rates of pay, including those of teachers, are now 'employer determined' on a 'take-it-or-leave-it' basis.

TEST YOURSELF 6.14
Describe one other condition of work that a trade union might bargain to improve besides those mentioned above.

The effect of introducing a trade union into a previously perfectly competitive labour market

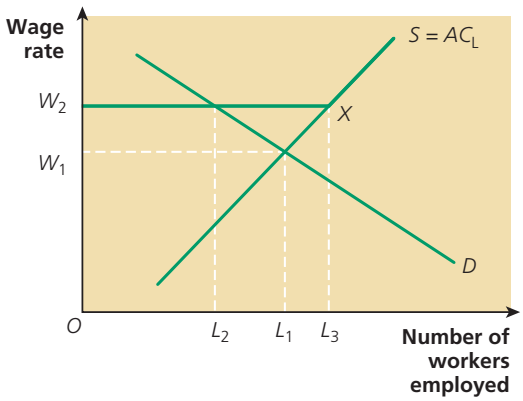


Figure 6.10 The effect of introducing a trade union into a previously competitive labour market

Figure 6.10 shows the possible effects of workers organising a trade union in a labour market that had previously been perfectly competitive. A first point to make is that the market would immediately become imperfectly competitive! It can be seen that, without a trade union, the competitive wage rate is W_1. After the workers join the union, through collective bargaining it negotiates a rise in the minimum wage rate acceptable to union members to W_2. Without the union, the market supply of labour curve is the upward-sloping line labelled $S = AC_L$. With the union, the market supply of labour curve is the kinked line W_2XS. For all sizes of labour force to the left of, or below, L_3, the supply curve of labour is horizontal or perfectly elastic, lying along the wage W_2 set by the trade union. If employers wish to hire a labour force larger than L_3 (and to the right of point X), a wage higher than W_2 has to be offered. Beyond L_3, the supply curve of labour slopes upwards because higher wage rates are needed to attract more workers.

At the wage level set by the union, employers only wish to hire L_2 workers. However, L_3 workers are willing to work at this wage rate. This means there is excess supply of labour and unemployment in the labour market. More workers wish to work than there are jobs available. The resulting unemployment is shown by the distance $L_3 - L_2$. However, the fall in employment compared to the competitive equilibrium is $L_1 - L_2$.

STUDY TIP
The effects shown in Figure 6.10 can also be used to explain the effect of a national minimum wage imposed above the wage rate that would be determined by market forces.

SYNOPTIC LINK
In Chapter 11, the type of unemployment shown in Figure 6.10 is called classical, or real-wage, unemployment. This is a useful link to be able to make when discussing the macroeconomic impacts of microeconomic issues.

The effect described above is sometimes used to justify the argument that any attempt by a union to raise wages must inevitably be at the expense of jobs, and that if unions are really interested in reducing unemployment, they should accept wage cuts. However, many economists — especially those of a Keynesian and left-of-centre persuasion — dispute this conclusion. They argue, first, that it is unrealistic to assume that conditions of demand for labour are unchanged. By agreeing to accept technical progress, by working with new capital equipment and new methods of organising work and by improving the skills of their members, a union can ensure (with the cooperation of management) that the *MRP* curve of labour shifts to the right. In these circumstances, increased productivity creates scope for both increased wages and increased employment. Higher wages may encourage firms to adopt improvements in productivity to pay for the higher wages but, on the other hand, some unions may resist the changes in working practices that lead to increased productivity.

Second, both wages and employment can rise when a union negotiates for higher wages in firms producing in an expanding goods market. In these conditions, increased demand for output creates increased demand for labour to produce the output. Indeed, rising real wages throughout the economy are likely to increase the aggregate demand for the output of all firms producing consumer goods because wages are the most important source of consumption expenditure in the economy.

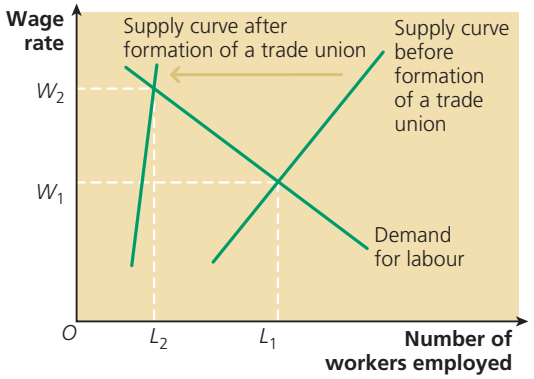


Figure 6.11 A trade union shifting the market supply curve of labour

So far, we have assumed that trade unions try to increase pay by preventing union members supplying labour at wage rates below the rate set by the union. Figure 6.11 illustrates a second way in which much the same result can be achieved. In this case, the trade union establishes a *closed shop*, which keeps non-union workers out of the labour market. The union-controlled entry barrier shifts the supply curve of labour to the left, and increases the inelasticity of the curve. A similar effect can be achieved by unions and professional associations that insist on long periods of training before the worker is formally qualified, during which time only very low wages are paid. Employment falls from L_1 to L_2, and the wage rate rises to W_2. The advantage to the union, and its members who are employed, is that there is no pool of unemployed workers who would be willing to work for lower wages — that is, there is not an excess supply of labour.

The effect of introducing a trade union into a monopsony labour market

The assertion that unions raise wages at the expense of jobs is heavily dependent on the assumption that, before the union was formed, the labour market was perfectly competitive (it is worth remembering that *no* real-world labour market can meet all the conditions of perfect competition). In the case of a monopsonistic labour market, it is possible for a union to raise *both* the wage rate and employment, even without the *MRP* curve shifting rightwards. This is illustrated in Figure 6.12. If the monopsonistic labour market is non-unionised, the equilibrium wage rate is W_1 and the level of employment is L_1.

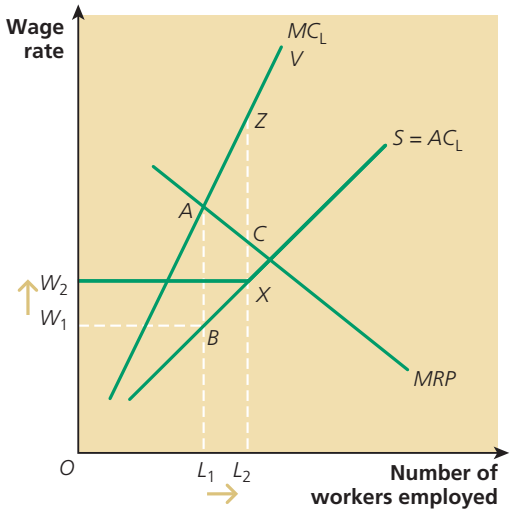


Figure 6.12 The effect of introducing a trade union into a monopsony labour market

The introduction of a trade union into a monopsonistic labour market has the same effect on the labour supply curve as would be the case if the labour market were to be perfectly competitive. In Figure 6.12, when the union sets the wage rate at W_2, the kinked line W_2XS becomes the labour supply curve (and also the AC_L curve). But in monopsony, W_2XS is *not* the marginal cost of labour curve. The MC_L curve is the double-kinked line W_2XZV. The double kink is explained in the following way.

Provided the monopsony employs a labour force smaller than or equal to L_2, the MC_L of employing an extra worker equals both the AC_L and the union-determined wage of W_2. But beyond L_2 and point X, the monopsony must offer a higher wage in order to persuade the members of an enlarged labour force to supply their labour. In this situation, with all the workers now being paid the higher wage rate, the marginal cost of labour increases significantly and there is a discontinuity, X to Z, in the MC_L curve; to the right of L_2 the MC_L curve lies above the AC_L curve. The upward-sloping line ZV drawn in Figure 6.12 shows the MC_L of increasing employment above the level L_2.

This means there is a discontinuity, shown by the vertical distance from point X to point Z, between the horizontal section of the MC_L curve (for levels of employment at or below L_2 and point X) and the upward-sloping section of the curve (ZV) to the right of point X. In the absence of a union, the level of employment is L_1, determined at point A (with point B determining the wage rate W_1). But when the union sets the wage rate at W_2, employment rises to L_2, which is the level of employment at which the MRP curve intersects the gap or discontinuity in the MC_L curve at point C, between X and Z. The union has managed to increase both the wage rate and the level of employment.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.4

Trade union action in the UK rail industry in 2022

The year 2022 saw the largest instance of industrial action in the UK since 1989, with trade unions in a variety of industries, such as university education, the legal profession, airlines and rail transport, voting for strike action.

Trade union action in the rail industry began in June 2022 as members of the National Union of Rail, Maritime and Transport Workers (RMT) went on strike over a dispute about pay and the threat of changes to working methods that would lead to some workers being made redundant. The RMT stated that Network Rail, a major employer of rail workers, was threatening to cut important safety-related jobs and increase working hours as part of a modernisation programme.

The initial strike followed the breakdown of negotiations between the RMT and 13 rail companies. Forty thousand rail workers went on strike as well as staff working on the London Underground, leading to significant disruption of the rail network. People in the UK were urged to make only essential journeys on strike days and many workers chose to work from home.

RMT union members voted to strike after their request for a 7% pay rise was rejected and a 3% offer made to the union members in response. This was at the time of a so-called 'cost-of-living crisis' in the UK. Three days of strike action were agreed, including on days when 90,000 students were due to sit an A-level Maths exam, with further strikes scheduled later in the summer. Network Rail refused at the time to make any further improvements to its pay offer and also threatened to make redundancies if the RMT did not halt its intentions to strike.

Critics of the rail strikes, including Network Rail, argued that the rail system needed to operate much more efficiently. This was because of a

significant change in commuter habits, such as reduced use of ticket offices and an increase in people working from home as a result of the Covid-19 pandemic. As an example, Network Rail referred to jobs that could be completed by one person currently being carried out by teams of up to nine people, and excessive walking times to staff canteens before and after breaks.

Follow-up questions

1 Explain what is meant by a trade union.
2 Why was the 'cost-of-living crisis' in 2022 an important factor in the trade union action of rail workers?
3 Research three trade unions (other than those mentioned here) which have been involved in industrial action in recent years. How successful have these unions been in negotiating improved pay and working conditions for their members?
4 Research what has been happening to the number of people belonging to a trade union over the last 30 years.

SECTION 6.5 SUMMARY

- A trade union is an association of workers formed to protect and promote the interests of its members.
- A major function of a union is to bargain with employers to improve wages and other conditions of work.
- We have assumed that a trade union is a monopoly supplier of labour, which is able to keep non-members out of the labour market and also to prevent its members from supplying labour at below the union wage rate.
- Free-market economists argue that, if a union raises the wage rate above the perfectly competitive wage rate, unemployment increases.
- In monopsony, a trade union may be able to increase both the wage rate and the level of employment towards the perfectly competitive levels.

6.6 The national minimum wage

The effects of a national minimum wage on labour markets

KEY TERM
national minimum wage a minimum wage or wage rate that must by law be paid to employees, though in many labour markets the wage rate paid by employers is above the national minimum wage.

The explanation in the previous section, of how a trade union can increase both the wage rate and the level of employment, can also be used to analyse the effect of a **national minimum wage** (NMW) set by the government. As we have already noted, if we assume a perfectly competitive labour market, Figure 6.10 can be adapted to show how a national minimum wage set above the market-clearing wage rate leads to a loss of jobs and unemployment. The same may be true if labour markets are highly competitive without being perfectly competitive. But if labour markets are monopsonistic, Figure 6.12, which likewise can be adapted to show how a national minimum wage can increase both the wage rate and the level of employment, might be more appropriate. However, this does not mean that the imposition of a minimum wage will *never* lead to a fall in employment. To repeat, some labour markets are very competitive even if they do not satisfy all the conditions of a perfectly competitive labour market.

QUANTITATIVE SKILLS 6.4

Worked example: plotting and interpreting data on a supply and demand graph in a labour market

The figures in Table 6.2 relate to the demand for, and supply of, gardeners at different hourly wage rates. The demand and supply schedules are in thousands of workers.

Table 6.2 The labour market for gardeners

Wage rate (£ per hour)	$Supply_1$	$Demand_1$	$Supply_2$
5	40	160	100
7	60	150	110
9	80	140	120
11	100	130	130
13	120	120	140
15	140	110	150
17	160	100	160

(a) Calculate the market wage rate and employment level for gardeners when the demand schedule is $Demand_1$ and the supply schedule is $Supply_1$.

The only hourly wage rate at which demand equals supply is £13. This is the equilibrium market wage rate, with the equilibrium level of employment being 120,000.

(b) The supply schedule now becomes $Supply_2$, but the demand schedule does not change. What happens to the equilibrium wage rate and level of employment?

The only hourly wage rate at which demand equals supply is £11. This is the equilibrium market wage rate, with the equilibrium level of employment being 130,000.

(c) Suppose the government now sets a national minimum wage at £13 per hour. How much unemployment would this create among gardeners?

Given that the supply schedule is still $Supply_2$, the national minimum wage rate of £13 has been set above the equilibrium market wage, which leads to 140,000 – 120,000 = 20,000 unemployment.

The advantages and disadvantages of a national minimum wage

A national minimum wage only really affects labour markets where the wage rate determined by the market is currently below the NMW rate. This, of course, includes markets where wages are reduced by monopsony power. When the national minimum wage rate is set below wage rates determined by free-market forces, then, just like any other minimum price law, it has no effect on the market.

SYNOPTIC LINK

See the coverage of government intervention in markets in Chapter 11. Many of the methods of government intervention can be applied to labour markets as well as markets for products.

Supporters of a national minimum wage argue that a 'wage floor' can be justified on the grounds that it reduces exploitation by employers of low-paid workers. As we have mentioned, many real-world labour markets, particularly those for low-paid unskilled workers, resemble the monopsony labour market shown earlier in Figure 6.9. In Figure 6.13 we have redrawn Figure 6.9, but with wage rate W_2 added to the diagram. This is the wage rate, determined at point C in the diagram, where the market demand for labour, depicted by the *MRP* curve, would equal the market supply of labour in a perfectly competitive labour market.

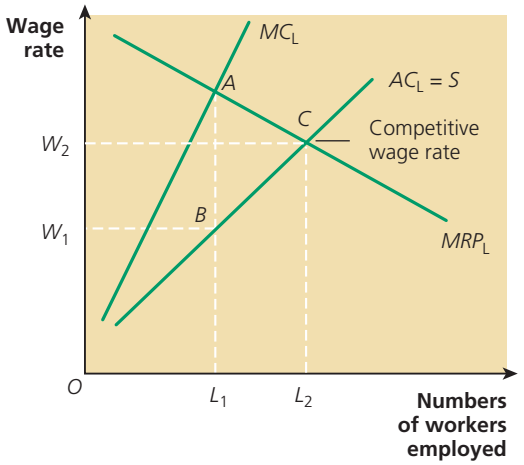


Figure 6.13 Comparing market wage rates in perfect competition and monopsony

Figure 6.13 illustrates the fact that in an imperfectly competitive labour market, the employer or employers use their market power to drive down the wage rate below the competitive rate of W_2 to the lower rate of W_1. Trade unions might, of course, create a situation in which the monopsony power of the buyers of labour (the employers) is matched by the monopoly power of the sellers of labour (the unions). This might restore the wage rate to W_2. But if this does not happen, a national minimum wage is deemed necessary, according to its supporters, to eliminate or reduce labour market exploitation of low-paid workers.

Exploitation of low-paid workers also increases poverty. Supporters of the national minimum wage believe that an NMW, set at a 'fair' level, should be a vital part of government policy which aims to prevent 'rogue' employers driving wage rates down to 'poverty rates'. (See Chapter 7 and the extension material on page 190.) By redistributing income from employers towards low-wage workers, a national minimum wage reduces income inequality as well as poverty.



Young workers are particularly likely to be affected by the national minimum wage

Most of the evidence collected since national minimum wages were first introduced suggests that 'wage floors' either slightly increase or slightly decrease employment. Either way, they do not result in mass unemployment. Recent studies generally conclude that, although there probably are unemployment effects, such effects are extremely slight. By raising the incomes of low-paid workers with only a slight impact on employment, according to this view, the benefits of national minimum wages far outweigh the costs.

Economists who believe in the virtues of a free-market economy, subject to the minimum possible level of government intervention, argue that the disadvantages of a national minimum wage exceed any possible advantages. In a briefing paper, 'The minimum wage: silver bullet or poisoned chalice?', published in 2014 by the Institute of Economic Affairs (IEA), a pro-free-market 'think tank' (a group that studies policy), Ryan Bourne and J.R. Shackleton argued that:

> There is evidence of reduced working hours in response to NMW increases. International research suggests many of the negative effects operate after a long time period through reducing new job creation.
>
> Employment impacts of increases in the minimum wage are likely to disproportionately affect the young, the unskilled, the long-term unemployed and those in lower productivity regions. There is evidence that minimum wages lead firms to replace lower-skilled and less experienced younger workers with older workers. The 18–24 year old unemployment rate has risen from 11.5% in April 1999 (when the NMW was introduced) to 17.9% today. And of those unemployed within this age group, the proportion out of work for more than 12 months has risen from 14.4% in 1999 to 31.8% in 2013.
>
> Claims employers are 'subsidised' by in-work benefits in the form of tax credits have some truth, but are exaggerated. Nearly a third of all tax credit recipient households do not have an adult in paid employment and a

further million households work fewer than 30 hours per week. Reforming tax credits would be a better means (than raising the NMW) of eliminating the degree to which tax credits subsidise the employers of the remaining full-time workers in receipt of credits.

Note, however, that the system of tax credits is set to end in 2024, to be replaced by Universal Credit.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.5

Should we have a national minimum wage?

A few years before the national minimum wage was introduced in the UK in April 1999 by a Labour government, in 1994 the National Bureau of Economic Research in the USA published a research article written by David Card and Alan B. Krueger. The article was titled 'Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania'.

Card and Krueger summarised their research as follows:

> On April 1, 1992 New Jersey's minimum wage increased from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and Pennsylvania before and after the rise in the minimum wage. Comparisons of the changes in wages, employment, and prices at stores in New Jersey relative to stores in Pennsylvania (where the minimum wage remained fixed at $4.25 per hour) yield simple estimates of the effect of the higher minimum wage.
>
> Our findings challenge the prediction that a rise in the minimum wage reduces employment. Relative to stores in Pennsylvania, fast food restaurants in New Jersey increased employment by 13 percent. We also compare employment growth at stores in New Jersey that were initially paying high wages (and were unaffected by the new law) to employment changes at lower-wage stores. Stores that were unaffected by the minimum wage had the same employment growth as stores in Pennsylvania, while stores that had to increase their wages increased their employment.

Card and Krueger's conclusions were seized on by British economists and politicians who supported the introduction of the national minimum wage. At the time, the British Conservative Party, then in opposition, opposed its introduction. However, by 2010 Conservative prime minister David Cameron, who initially opposed the national minimum wage, publicly stated that it 'turned out much better than many people expected'. This cautious acceptance of the minimum wage among policy-makers is now reflected throughout developed economies where few countries remain without a statutory wage floor.

However, many economists continue to oppose a national minimum wage, arguing that it distorts free markets and creates unemployment. In 2011 Karthik Reddy, in an article commissioned by the IEA, concluded that:

> Far from having positive effects for the working poor, the implementation of the minimum wage in the UK is likely to have had a negative effect on employment. The minimum wage has been routinely raised at a rate greater than inflation, resulting in a tremendous increase in the real value of the minimum wage. This rise is particularly problematic for workers during times of recession. The present government could alleviate the difficulties of the minimum wage by resisting the urge to raise it further and gradually allow inflation to reduce its real value. A slow death of the wage floor would stimulate employment and permit Britain to regain a competitive edge.

Follow-up questions

1. Using appropriate diagrams, analyse and compare the effect of a national minimum wage rate in competitive and monopsonistic labour markets.
2. Are Card and Krueger's conclusions compatible with Figure 6.12? Justify your answer.
3. Distinguish between the national minimum wage (NMW) and the national living wage (NLW).
4. At what rates is the national minimum wage set at the time you are reading this case study?

EXTENSION MATERIAL

The national living wage replaces the national minimum wage

Students have often confused the official *national minimum wage* (NMW) with the concept of a *living wage*. The NMW was introduced by a Labour government in 1999, and then continued by the Coalition government elected in 2010. The living wage, however, is often defined as the minimum income needed for a worker to meet their basic needs such as food, clothing and housing. The Conservative government which replaced the Coalition government in 2015 bowed to popular pressure and replaced the national minimum wage with a national living wage (NLW) in 2016, but only for workers aged 25 and older.

Both the NMW and latterly the NLW have been increased in most years by an amount more or less in line with the inflation rate measured by the consumer prices index (CPI). The government decides their levels after taking advice from the Low Pay Commission.

When it was first introduced, the NMW rate was £3.60 per hour (£3.00 for 18- to 21-year-olds). In July 2015, the adult NMW rate was £6.50 per hour. The most recent increase in the adult NMW for workers aged 21 to 22 took place in April 2022, with the rate rising to £9.18 per hour for those workers. At the same time, the NLW was increased to 9.50 per hour. At the time of its introduction, the NLW was set at £7.20 per hour, for both full-time and part-time workers. The various levels for both the NLW and the NMW in 2022 were as shown in Table 6.3.

Table 6.3 National living wage and national minimum wage rates, April 2022

Age group	Hourly rate (£)
National living wage (23+)	9.50
National minimum wage (21–22)	9.18
National minimum wage (18–20)	6.83
National minimum wage (16–17)	4.81

Table 6.4 shows the actual level of the NMW and the NLW for workers aged 25 and over between 2017 and 2022 (23 and over from 2021). It would be a good idea for you to check the current levels of both the NMW and the NLW at the time when you read this.

Table 6.4 National minimum wage and national living age rates for workers aged 25 and over (23 and over from 2021), 2017–22

	Hourly rate (£)					
	2017	2018	2019	2020	2021	2022
National minimum wage	7.05	7.38	7.57	7.76	8.91	9.50
National living wage	7.50	7.83	8.21	8.72	8.91	9.50

TEST YOURSELF 6.15

Distinguish between the national minimum wage (NMW) and the national living wage (NLW).

QUANTITATIVE SKILLS 6.5

Worked example: nominal and real values

In economics, nominal values are those which don't take inflation into account while real values are those which have had the effects of inflation taken into account. Between 2010 and September 2017 most workers employed in the UK public sector saw their nominal wage rates increase by 1% per year. This pay cap was largely abolished in July 2018, though the government said that any pay increases above 1% a year would have to be self-financing. The rate of inflation, measured by changes in the consumer prices index (CPI) in the 12 months before July in each year, was 4.4% in 2011, 2.6% in 2012, 2.8% in 2013, 1.6% in 2014, 0.1% in 2015, 0.6% in 2016, 2.6% in 2017 and 2.3% in 2018.

How did real wages paid to public-sector workers experiencing a 1% annual pay cap change between July 2011 and July 2018? Briefly comment on your answers.

The change in the real wage rate in each year equals the change in the nominal wage rate minus the rate of inflation. This means that, according to the data, the change in the real wage rate was –3.4% in 2011, –1.6% in 2012, –1.8% in 2013, –0.6% in 2014, +0.9% in 2015, +0.4% in 2016, –1.6% in 2017 and –1.3% in 2018.

Because the CPI inflation rate was higher than the 1% pay cap in all but two of the years covered by the pay cap, according to the data, public-sector real wages fell in most of these years; 2015 and 2016 were the only years in which public-sector real wages increased slightly.

SECTION 6.6 SUMMARY

- The explanation of how a trade union can increase both the wage rate and the level of employment can also be used to analyse the effect of a national minimum wage (NMW) set by the government.
- Opponents of a national minimum wage argue that it increases unemployment.
- Supporters of a national minimum wage argue that a 'wage floor' can be justified on the grounds that it reduces exploitation by employers of low-paid workers.
- From 2016 onward, for most adults the national living wage (NLW) replaced the national minimum wage (NMW).

6.7 Discrimination in the labour market

The nature of wage discrimination

KEY TERM
wage discrimination paying different workers different wage rates for doing the same job.

In imperfectly competitive labour markets, employers often possess sufficient market power to reduce the total wage bill through undertaking **wage discrimination**, which involves paying different workers different wages for doing the same job. Figure 6.14 illustrates the effect of wage discrimination introduced into a previously competitive labour market. We will assume that in the absence of wage discrimination, all workers are paid the same wage, W_1, determined by supply and demand. Employers' total wage costs are shown by the rectangle OW_1AL_1.

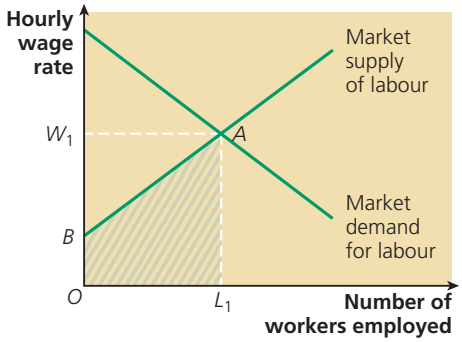


Figure 6.14 Wage discrimination

But if, instead of paying W_1 to all workers, employers pay each worker the minimum he or she is prepared to work for, the total wage bill falls to equal the shaded area $OBAL_1$. Employers thus gain at the expense of workers, which is one reason why firms pay, and trade unions resist, discriminatory wages whenever possible.

Wage discrimination can also occur when a group of workers is systematically paid a wage that is below their *MRP*. This occurs when employers successfully exploit their monopsony power in the labour market, as illustrated in Figures 6.12 and 6.13.

STUDY TIP

You should understand the similarity between *wage discrimination* in the labour market where a firm hires its labour and *price discrimination* in the goods market where the firm sells its output.

EXTENSION MATERIAL

Bringing together wage discrimination and price discrimination

In the case of perfect price discrimination (often called first-degree price discrimination), a firm charges different prices to different customers for the same good, with the same marginal cost of production, so that each customer pays the maximum price he or she is prepared to pay. The entire consumer surplus that customers would otherwise enjoy is transferred to the firm, enlarging the firm's profit.

With perfect wage discrimination, the firm pays each worker the minimum wage the worker is prepared to accept, without transferring their employment elsewhere. The part of the wage that workers would otherwise get is transferred to the firm, once again boosting profit.

In real life, wage discrimination is most likely to occur in imperfectly competitive labour markets. From an imperfectly competitive firm's point of view, the best possible outcome is simultaneous price discrimination in the goods market where it sells its output and wage discrimination in the labour market where it hires its workers. Profit is boosted from two directions at once.

Can you think of reasons why such simultaneous exploitation seldom takes place?

A condition necessary for wage discrimination

Successful wage discrimination, through which employers pay different wage rates to different groups of workers, requires employers to be able to identify and separate different groups of workers supplying the same type of labour. This is possible when workers differ in their knowledge of the labour market and in their ability to shop around among employers.

Types of wage discrimination

So far we have defined wage discrimination in terms of employers paying workers the minimum they are prepared to work for, without focusing on issues of gender, ethnicity and other aspects of discrimination. There are many kinds of discrimination. We will look at gender discrimination as an example, but it is worth researching how other groups are discriminated against, too.

STUDY TIP

Be aware that there are a number of different types of discrimination operating in different labour markets. These include racial, religious, gender, sexuality and disability discrimination.

Gender discrimination

Along with race discrimination, gender discrimination is significant. In recent years, although women have accounted for an increasing share of total employment in the UK, women's pay often continues to be lower than men's pay, despite the fact that equal pay legislation has been in place since 1972. The pay gap between men and women is at risk of widening for the first time on record, a leading pay equality campaign group has warned. In November

2017, the Fawcett Society said that progress had stalled in closing the gender pay gap, which stood at 14.1% — the same as it was in 2015 and 2016. The pressure group stated that if the gap continued to close at the rate it had since 2012, it would not reach zero until a century later in 2117. The gap was higher in the private sector, at 17.1%, but it had fallen by 4.3 percentage points since 2011. In the public sector it had stayed flat at just above 14%. Fawcett research published earlier in 2017 found that the mean aggregate pay gap for Pakistani and Bangladeshi women was 26% and for black African women it was 24%.

Women are almost twice as likely (1.8 times more likely) to receive the lowest pay — with 221,000 women earning less than the statutory minimum wage, 100,000 more women than men.

A second survey, the 2017 Gender Salary Survey, undertaken by the Chartered Management Institute (CMI) found that the gender pay gap stood at 26.8%, with male managers on average out-earning female peers by £11,606 a year — £3,000 more than previously thought.

There are several reasons why women earn less than men:

- Women work predominantly in low-paid industries and occupations. Within many occupational groups, women are paid less than men. This is often because women are under-represented in the higher-paid posts within an occupation, though it is also the case that many women are paid less for doing the same job.
- Discrimination against women in labour markets may contribute to these circumstances. In addition, women are disproportionately highly represented in industries where the average size of firm and plant is small. These industries tend to pay lower wages and offer fewer promotion prospects than large firms and large industries. Such industries are also seldom unionised. Indeed, within all industries, women workers traditionally have been less unionised than men.
- This relates to another reason why women earn less than men: on average, their attachment to the labour force is weaker. Each year of work experience raises the pay of both men and women by an average 3%. Yet if women leave the labour force, usually to look after young children, their potential pay falls by 3% for each year involved. For example, consider a man and woman who enter employment with equal potential and after eight years the woman leaves the workforce in order to raise a family. If she re-enters the labour force eight years later, she will be 16 years, in pay terms, behind the man.
- The higher labour turnover of women also imposes costs on the employer — for example, the costs of training replacement workers. This may reduce the incentive for employers to train female workers. Similarly, some women may have less incentive to spend time and money on their own education and training if they expect the benefits that they will eventually receive to be less than the costs initially incurred.

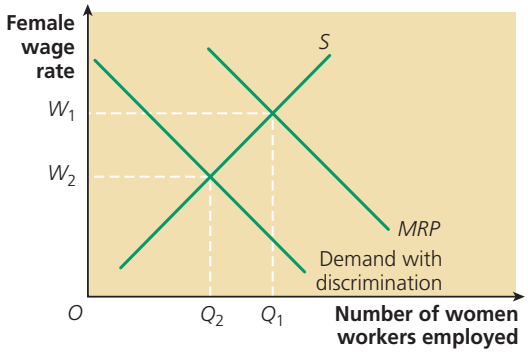


Figure 6.15 How gender discrimination may affect a labour market

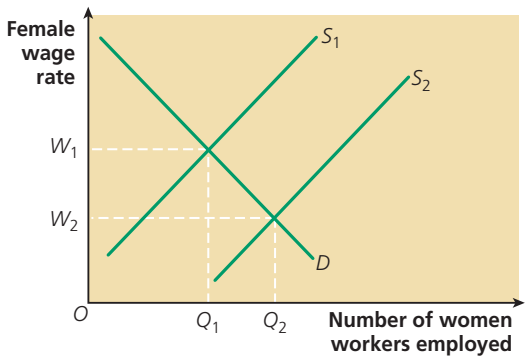


Figure 6.16 The effect of gender discrimination in one market upon female employment and wage rates in other labour markets

Figures 6.15 and 6.16 illustrate how gender discrimination may affect labour markets. In Figure 6.15 the demand curve for female workers, shown as 'Demand with discrimination', lies to the left of their marginal revenue product curve. This results in Q_2 women being employed, with each being paid an hourly wage rate of W_2. Without gender discrimination, female employment would be Q_1, with the wage rate at W_1.

TEST YOURSELF 6.16
Name three ways in which wage discrimination can occur in a labour market.

If gender discrimination takes place in one labour market, other labour markets in which no discrimination is taking place will also be affected. In these markets, as Figure 6.16 shows, the supply curve of female workers increases as a direct result of discrimination taking place in the first labour market. Female employment increases from Q_1 to Q_2, and the female wage rate falls from W_1 to W_2.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.6

Glass ceiling for women replaced by reinforced concrete as progress stalls

The so-called glass ceiling that prevents women progressing in the workplace is more like 'reinforced concrete', Nicola Brewer, the chief executive of the Equality and Human Rights Commission (EHRC), has warned:

> We always speak of a glass ceiling. These figures reveal that in some cases it appears to be made of reinforced concrete. We need radical change to support those who are doing great work and help those who want to work better and release talent. Young women's aspiration is in danger of giving way to frustration. Many of them are now excelling at school and achieving great things in higher education. And they are keen to balance a family with a rewarding career. But workplaces forged in an era of stay-at-home mums and breadwinner dads are putting too many barriers in the way — resulting in an avoidable loss of talent at the top.

There are fewer women MPs in Westminster. However, 208 female MPs were elected during the 2017 General Election — a record high and 32% of all MPs. This was up from 191 in the 2015 election and the highest proportion of any UK election to date. The UK government's target of 25% female membership by 2015 has now been reached.

The EHRC has likened women's progress to a snail's pace. According to the commission, a snail could crawl:

- nine times round the M25 in the 55 years it will take women to achieve equality in the senior judiciary
- from Land's End to John O'Groats and halfway back again in the 73 years it will take for equal numbers of women to become directors of FTSE 100 companies (FTSE 100 is the hundred companies with the highest market capitalisation whose share prices are quoted on the London stock exchange)
- the entire length of the Great Wall of China in 212 years, only slightly longer than the 200 years it will take for women to be equally represented in Parliament.



Poppy Gustafsson, chief executive officer of Darktrace PLC, a British-American information technology company

The commission's report argued that its findings are not just a 'women's issue' but a powerful symptom of a wider failure. The report asked in what other ways old-fashioned, inflexible ways of working are preventing the UK from tapping into talent — whether that of women or other under-represented groups such as disabled people, ethnic minorities and those with caring responsibilities. It argued that the country cannot afford to go on marginalising or rejecting talented people who fail to fit into traditional work patterns.

Follow-up questions

1. What is meant by a 'glass ceiling'?
2. Do 'glass ceilings' adversely affect other groups in society, such as ethnic and religious minorities and old people, or does discrimination take different forms against these groups of people?
3. Research up-to-date statistics on women's representation in Parliament since the 2015 general election, and whether the government's 25% target for female company directors has been met.

APPLICATION OF ECONOMICS IN THE REAL WORLD 6.7

A real-world example of wage discrimination: Ford car workers in 1968 and after

In 2010 the film *Made in Dagenham* was released, to be followed in 2014 by a West End musical with the same name. The two events drew to popular attention the important story of how UK law was changed by the action of women employed at Ford's Dagenham car plant in the East End of London in 1968 and in the years thereafter. Here is part of the real story of *Made in Dagenham*.

Back in 1968, the idea of a male 'breadwinner' bringing in a 'family wage' legitimised lower pay for women. However, in that year, women sewing machinists in Ford's car plant in Dagenham took a stand for equal pay in a strike that stopped production for three weeks. The Ford strike, which captured the attention of Barbara Castle, the secretary of state for labour, led to the 1970 Equal Pay Act.

The women machinists were unhappy that their jobs had been downgraded from Category C (skilled production jobs) to Category B (unskilled production jobs), and that they would also be paid 15% less than men in grade B for doing work of a similar skill level. Three weeks later, following Barbara Castle's involvement, their rate of pay was increased to 8% below the men's rate in grade B. However, a grade C rating was only achieved in 1984, following a second strike.

Follow-up questions

1. Find out whether wage differences between male and female workers are greater in the private sector than in the public sector.
2. What is the difference between skilled and unskilled work? Give an example of each of these types of work.
3. Research details about the extent to which women workers remain underpaid in the UK today.
4. As part of the process of globalisation, car production has moved away from Europe and North America to countries such as China and South Korea. What effect worldwide do you think this might have had on female pay?

SECTION 6.7 SUMMARY

- In monopsony, a trade union may be able to increase both the wage rate and the level of employment towards the perfectly competitive levels. A national minimum wage can have the same effect.
- From 2016 onwards, for most adults the national living wage (NLW) replaced the national minimum wage (NMW).
- Perfect wage discrimination occurs when workers are paid the minimum wage rates they are prepared to accept.
- Gender discrimination means that women often earn less than men.

Questions

1 Why might it be said that roles are reversed between the goods market and the labour market in an economy?

2 Explain three factors that are significant in determining a firm's demand for labour.

3 Explain the shape of a market supply curve of labour.

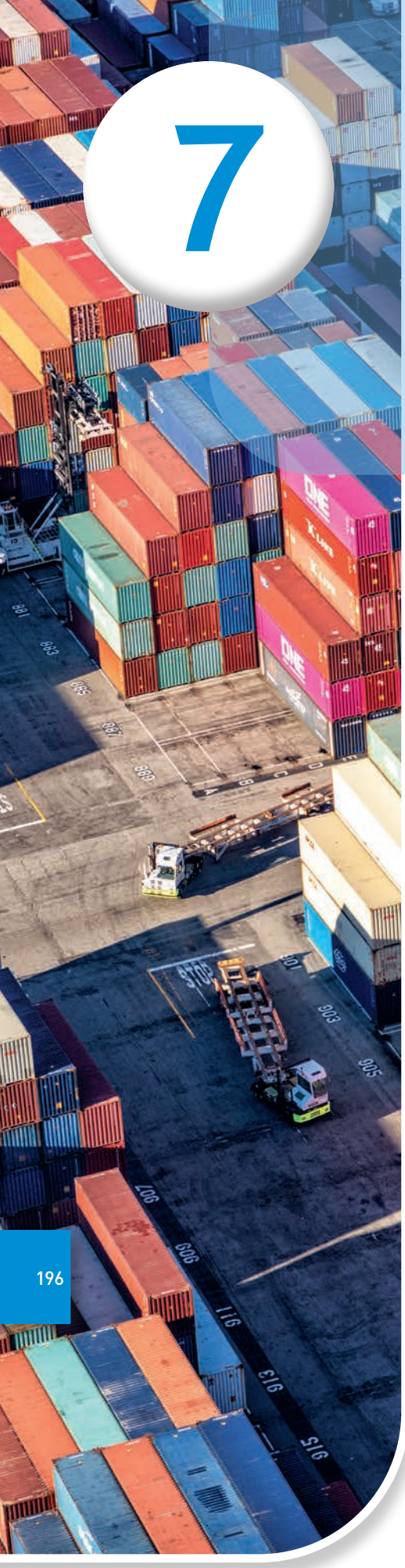
4 What is a perfectly competitive labour market?

5 Distinguish between the elasticity of demand for labour and the elasticity of supply of labour.

6 How is the wage rate determined in a monopsony labour market?

7 Evaluate the view that when trade unions raise wages they inevitably reduce levels of employment.

8 Assess whether gender wage discrimination has been significantly reduced in the UK in recent years.



7

The distribution of income and wealth: poverty and inequality

The way the labour market operates, which was analysed in Chapter 6, contributes significantly to inequalities in the distribution of income. This chapter builds on this analysis, covering also wealth inequalities and causes of poverty, which result partially from income and wealth inequalities.

Unregulated market forces tend to produce highly unequal distributions of income and wealth in many countries, which many economists believe to be an important form of the market failures which will be investigated in Chapter 8. However, a minority of economists disagree. Extreme pro-free-market economists sometimes argue that people who end up poor deserve to be poor. According to this view, the market does not fail; it simply creates incentives that cause people to generate income and wealth which end up benefiting most of the population. They also argue that attempts by governments to redistribute income and wealth from the rich to the poor usually end up, through the distortion of personal incentives, in government failure which harms national economic performance. Nevertheless, the mainstream economic view is that some form of intervention to moderate the high levels of inequality that would occur in a completely free-market economy is desirable. The arguments are about the *extent* to which governments should intervene to reduce inequality, and the methods of intervention.

For households at the bottom of the income distribution, severe income inequalities are an important cause of poverty (although, as mentioned above, pro-free-market economists argue that the resulting incentives provide a route out of poverty). However, poverty has causes other than those associated with income inequality, and we will examine these later in the chapter.

While this chapter explains inequalities and poverty in a microeconomic way, they reappear in Part 2 of the book when, for example, investigating the role of fiscal policy in both creating and trying to solve inequalities and poverty.

LEARNING OBJECTIVES

These are to understand:

- the nature and causes of inequalities in the distribution of income and wealth in the UK
- the difference between equality and equity in relation to the distribution of income and wealth
- how government policies such as progressive taxation, transfers and the national minimum wage attempt to make the distribution of income more equal
- the meaning of poverty
- the difference between absolute poverty and relative poverty
- the main causes and effects of poverty in the UK
- how progressive taxation relates to fiscal drag and to the poverty and unemployment traps.

7.1 The distribution of income and wealth

The difference between income and wealth

Together with other members of the general public, economics students often confuse the two words, **income** and **wealth**. The main difference between them is that income is a *flow* whereas wealth is a *stock*. Personal wealth is the stock, or historical accumulation, of everything you own that has value. For those households with at least one working member, wage and salary payments into the household's bank account are an inward flow of income, whereas, for an owner-occupying household, the house they live in, minus any outstanding **mortgage** debt, usually forms a large part of their household wealth.

To recap, your personal income is the *flow* of money you receive hourly, weekly, monthly or annually, some of which (the part that you *save*) can add to your personal wealth. As we will see, this is one of the links between income and wealth. A second link operates in the opposite direction — the wealthier you are, the more investment income or unearned income you are likely to earn, which adds to your total income. Indeed, the rich benefit from a *virtuous* circle: wealth increases income, which allows the wealthy to save, and saving adds to wealth, and so on. By contrast, many of the poor suffer a *vicious* circle: low income means the poor have to borrow, borrowing adds to personal debt, income is then spent on debt repayment, consumption falls, and any wealth the poor possess disappears.

The **distribution of income** measures how personal or household income is distributed among different income groups in society, such as between rich and poor. The term is also used to measure other forms of distribution — for example, between people living in different parts or regions of the economy, between different generations (such as the old and the young), and between men and women. Measures of income distribution can also be extended to the international distribution of income between countries — with countries at different stages of economic development having not only different total levels of income, but also different distributions between rich and poor — and also between different regions, age groups and gender.

Likewise, the **distribution of wealth** measures how personal or household wealth is distributed among different groups in society. Again, the distribution between rich and poor (and the intermediate groups which separate rich and poor) is most often considered, though economists sometimes focus on differences in regional, age group and gender distributions of wealth.

KEY TERMS

income personal or household income is the flow of money a person or household receives in a particular time period.

wealth personal wealth is the stock of everything that a person or household owns at a particular point in time which has value.

mortgage a long-term loan usually used to buy a house. The loan is secured against the value of the house.

distribution of income how income is divided between rich and poor, or between different groups in society, e.g. on a regional, age or gender basis.

distribution of wealth how wealth is divided between rich and poor, or between different groups in society, e.g. on a regional, age or gender basis.

SYNOPTIC LINK

Chapter 10 explains national income and national wealth in a macroeconomic context.

TEST YOURSELF 7.1

Which of the following are forms of income rather than forms of wealth?

(a) A national savings certificate; (b) housing benefit; (c) an inheritance; (d) interest paid to the owner of a government bond; (e) prize money won; (f) a dividend paid to a shareholder (a share of a company's profit given to someone who has made a financial investment in the company).

QUANTITATIVE SKILLS 7.1

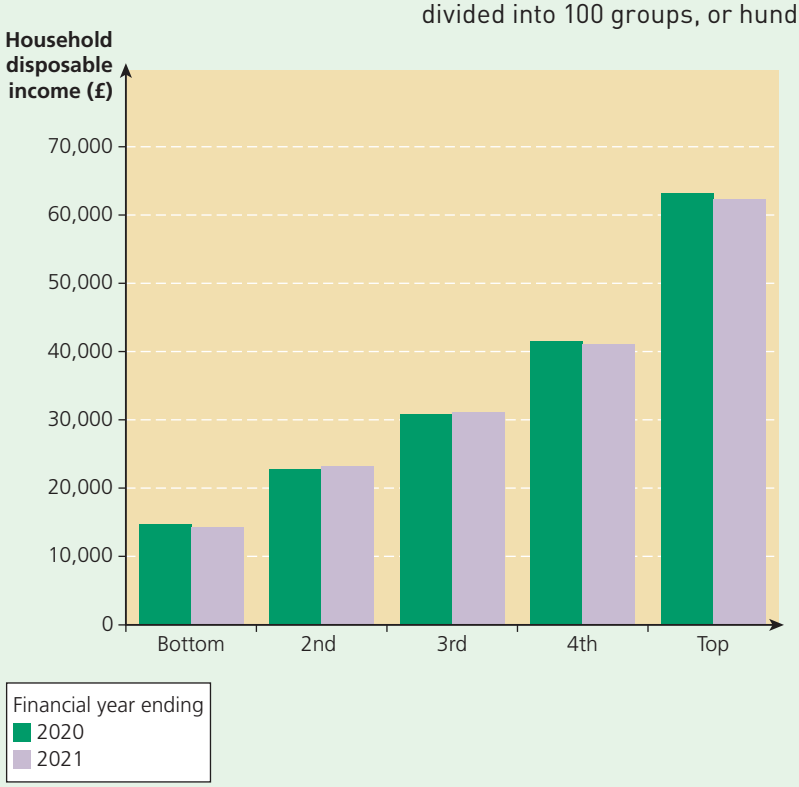
Worked example: quintiles and quantiles

Figure 7.1 shows quintile data for the distribution of income in the UK.

Explain what is meant by a quintile. Describe two other quantiles that economists often use when they present data.

Note that the Annex to the A-level specification states that you must understand and use the terms 'mean' and 'median' and relevant quantiles.

As the second part of the question implies, a quintile is an example of a quantile. A quantile is a portion of a total amount divided into sub-groups. In Figure 7.1, total income has been divided into five equal-sized groups, or fifths. These are quintiles. Other commonly used quantiles are quartiles (the total is divided into four groups, or quarters); deciles (the total is divided into ten groups, or tenths); vigintiles (the total is divided into 20 groups, or twentieths); and finally, percentiles (the total is divided into 100 groups, or hundredths).



Source: ONS

Figure 7.1 Household disposable income in the UK, by quintile, financial years ending 2020 and 2021

TEST YOURSELF 7.2

What is a decile?

EXTENSION MATERIAL

How the World Bank classifies countries according to income per capita

The World Bank classifies world economies into four groups: low income; lower-middle income; upper-middle income; and high income. Income is measured using gross national income (GNI) per capita, in US dollars. GNI is the total value of goods and services produced by an economy in a year, including net incomes earned by factors of production overseas. Per capita means divided by the size of the country's population.

In July 2021 the bank updated the thresholds for each classification as shown in Table 7.1.

Table 7.1 The World Bank's 2021 thresholds for classification of countries by annual income

Threshold	GNI/capita (current US$)
Low income	<1,046
Lower-middle income	1,046–4,095
Upper-middle income	4,096–12,695
High income	>12,695

The UK is, of course, a high-income country with GNI per capita (according to the World Bank) of $45,380 in 2021. Turkey is an upper-middle-income country (GNI per capita $9,830 in 2021); Bolivia is a lower-middle-income country (GNI per capita of $3,360 in 2021); while in sub-Saharan Africa the Democratic Republic of Congo is a low-income country (GNI per capita of $580 in 2021).

SYNOPTIC LINK

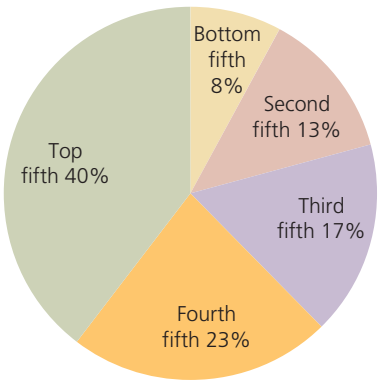
Chapter 15 provides further details of income groups in the context of the United Nations Human Development Index.

Factors which influence the distribution of income and wealth

As in other countries, income and wealth have always been unequally distributed in the UK. Even when economic growth creates full employment, the incomes of the rich tend to increase faster than those of the poor. For this reason, fast economic growth may actually widen income differences, though those receiving little money can still end up *absolutely* better off.

Factors influencing the distribution of income

Figure 7.2 shows that in 2018 the top fifth of UK households received 40% of total household income, whereas the bottom fifth received only 8%. The data relate to income after taxes have been deducted and welfare benefits received. A large number of factors influence the distribution of income, as described below.



Note: the figures do not add up to 100% due to rounding.
Source: The Equality Trust
Figure 7.2 How income was shared in the UK in 2018

Factors of production

An important factor is the distribution of national income between the different factors of production: land, labour, capital and entrepreneurs. The factors of production were introduced and explained in Chapter 1. The owners of large land holdings, such as country estates or large parts of the West End of London, receive large incomes in the form of rent. This puts large landowners in the top income quintiles shown in Figure 7.1. The share of national income of landlords and owners of capital has grown, whereas labour's share has fallen. After being largely stable in many countries for decades, the share of national income paid to workers has been falling since the 1980s. The April 2017 edition of the *World Economic Outlook* published by the International Monetary Fund (IMF) found that this trend is driven by rapid progress in technology and global integration.

Across the world, labour's share of worldwide incomes was 62% in the early 2000s, down from over 66% in the early 1990s. One reason for this is that wages have not been rising as fast as productivity and hence labour's share of national income has fallen. Many of the owners of capital, including

entrepreneurs, receive their income in the form of dividend payments on the shares they own. Profits and entrepreneurial incomes have risen at the expense of wages and salaries, although the salaries received by top business executives have also grown much faster than the wages and salaries of other workers.

The distinction between earned and unearned income

Earned income includes wages, salaries, other forms of employee compensation, and self-employment income. Unearned income is income derived from sources other than employment, including interest and investment income (such as dividend income paid to shareholders). The various determinants of wages discussed in Chapter 6 can be used to explain differences in earned incomes. By contrast, differences in unearned income depend primarily on inequalities in the distribution of wealth.

Wage and salary differentials

Within the labour market, the difference between the wages and salaries of those at the top and those at the bottom have widened. To an extent this can be explained by standard supply and demand theory.

SYNOPTIC LINK

The factors which determine labour demand and supply are covered in detail in Chapter 6, especially in sections 6.1 and 6.2. You should familiarise yourself with these concepts in order to understand wage and salary differentials between occupations.

Figure 7.3(a) shows a high hourly wage rate in the market for airline pilots, while Figure 7.3(b) shows a much lower wage rate in the market for aircraft cleaners. Two of the main factors contributing to this difference in hourly wage rates are the differences in labour productivity, which determine the relative positions of the demand curves for pilots and for cleaners, and the different slopes and positions of the supply curves.

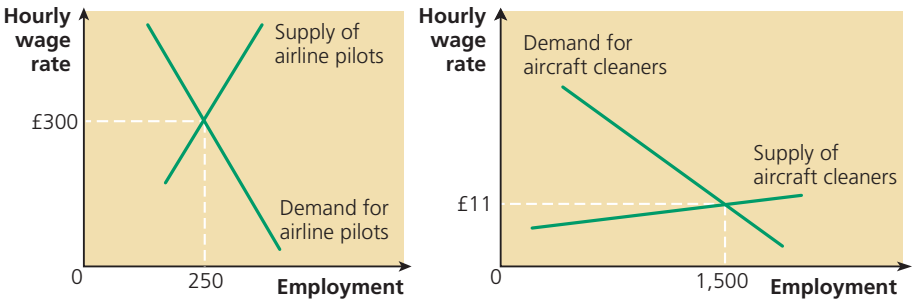


Figure 7.3 Using supply and demand to explain differences in rates of pay in different jobs

The labour productivity of an airline pilot is significantly higher than that of an aircraft cleaner. The demand for pilots is less wage elastic than the demand for cleaners. Similarly, because of factors such as long training periods, the supply of airline pilots is lower and more inelastic with respect to changes in their salaries. By contrast, the supply of airline cleaners is higher at any given wage and more wage-rate elastic. The cleaners are unskilled and do not require much training. As a result, an increase in the wage paid to airline cleaners greatly increases the number of workers willing to clean aircraft. Because of

these factors, and others we have not considered, pilots' salaries end up being much higher than the wages received by the workers who clean their aircraft. In Figure 7.3, pilots are paid £300 an hour, whereas cleaners earn £11 an hour, which is just above the 2023 national living wage.

Globalisation and the international migration of workers

Globalisation and the international migration of workers have also widened the differences between the wages and salaries paid to different groups of workers. In the UK, low-paid workers are in competition both with incoming migrants from poorer parts of the world, and with overseas-based workers employed in developing economies, to which UK employers have outsourced jobs. A call-centre worker, serving the UK market but employed in India, earns a much lower wage than a similar worker employed in the UK. International competition has led to falling wages in UK labour markets where workers are competing for jobs against similar workers in other countries.



UK outsourcing of jobs to call centres in developing economies has led to falling wages for some British workers

Similar competition exists at the high end of labour markets too, but in this case, well-paid UK business executives often succeed in raising their pay on the grounds that their employers must match the 'rate for the job' established in richer countries such as the USA. If their pay does not rise, they argue, they will move to better-paid jobs in other countries.

SYNOPTIC LINK

See Chapters 1 and 6 for more information on the factors of production and Chapter 14 for discussion of globalisation.

APPLICATION OF ECONOMICS IN THE REAL WORLD 7.1

Causes of declining labour share of income

According to an IMF blog published on 12 April 2017, from which this Application of Economics in the Real World has been taken, labour's share of income declines when wages grow more slowly than productivity, or the amount of output per hour of work. The result is that a growing fraction of productivity gains has been going to capital. And since capital tends to be concentrated in the upper ends of the income distribution, falling labour income shares are likely to increase income inequality.

In advanced economies, labour income shares began trending down in the 1980s. They reached their lowest level of the past half-century just prior to the 2008 global financial crisis and have not recovered materially since. Labour income shares are almost 4 percentage points lower now than they were in 1970.

Despite more limited data, it is also the case that labour shares have declined in emerging markets and developing economies since the early 1990s. This is especially true for the larger economies in this group. In China, for example, despite impressive gains in poverty reduction over the past two decades, labour shares still fell by almost 3 percentage points.

Indeed, as growth remains below average in many countries, an increasing recognition that the gains from growth have not been broadly shared has strengthened a backlash against economic integration and bolstered support for inward-looking policies. This is especially the case in several advanced economies. In advanced economies, about half of the decline in labour shares can be traced to the impact of technology. The decline was driven by a combination of rapid progress in information and telecommunication technology, and a high proportion of occupations that could be easily be automated.

Follow-up questions

1 Distinguish between labour and capital.
2 What is meant by an inward-looking policy?
3 The account above states that 'labour's share of income declines when wages grow more slowly than productivity'. Why should this be so?
4 In 2014 the French economist Thomas Piketty published an important book on the distribution of income between labour and capital. Access Piketty's lectures on YouTube: for example, 'New thoughts on capital in the twenty-first century'. What are Piketty's views on the main causes of inequality?

STUDY TIP

Make sure you can apply the theory of the labour market explained in Chapter 6 to the factors determining the distributions of income and wealth. Some of the key reasons for differences in income and wealth are due to differences in the demand for, and supply of, labour.

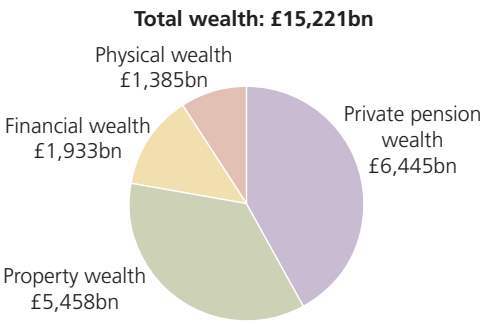
Inequality in the distribution of wealth

While Figures 7.1 and 7.2 both show that the distribution of income in the UK is unequal, the distribution of wealth is significantly more unequal.

According to the Resolution Foundation, a UK think tank whose aim is to improve the standard of living of low- and middle-income families, the spreading of property wealth across the UK in the mid-1990s and mid-2000s contributed to a decade-long fall in wealth inequality. However, that trend has now gone into reverse.

The UK's record £15.2 trillion of wealth in 2020 was distributed far less equally than earnings or household income. Currently, one in ten adults own around half of the nation's wealth, with the top 1% owning 14% of UK wealth. By contrast, 15% of adults have no or negative wealth.

The Office for National Statistics (ONS) defines total net wealth as the sum of four components: net property wealth, physical wealth, net financial wealth and private pension wealth, although it does not include rights to state pensions. These are shown in Figure 7.4 for 2020.



Source: ONS

Figure 7.4 Breakdown of the types of wealth owned by UK households, 2020

STUDY TIP

These days, UK data, particularly in macroeconomics but also in microeconomic topics such as the distributions of income and wealth, are increasingly expressed in trillions of pounds. Make sure you understand the difference between a trillion, a billion and a million.

As we mentioned earlier, wealth is a 'stock' concept rather than a 'flow'; it is measured at a particular point in time. Income, which refers to the flow of resources received over a period of time, allows wealth to be accumulated, but similarly, wealth is capable of producing flows of income either in the present or — as in the case of pension wealth — in the future.

QUANTITATIVE SKILLS 7.2

Worked example: calculating percentages to 2 decimal places

Figure 7.4 shows the relative contribution of each of the four wealth components to aggregate total wealth in the UK in 2020. Calculate to 2 decimal places the percentage contribution of each of the four components of wealth to total wealth in 2020.

In 2020, the two components making the largest contribution to aggregate total wealth were private pension wealth and net property wealth (accounting for 42.34% and 35.86% respectively). Financial wealth made up 12.73% of the total wealth in 2020 and physical wealth (other than property wealth, e.g. cars) made the smallest contribution of the four components (9.10%).

Factors influencing the distribution of wealth

The factors influencing the distribution of wealth are closely linked to the factors which influence the distribution of income. Indeed, as we have already mentioned, high incomes lead both to a large amount of saving taking place, and to spending on items such as houses, both of which add to household wealth. By contrast, households living on small incomes can afford neither to save nor to purchase houses, which appreciate rather than depreciate in value.

Some of the other factors that influence the distribution of wealth are described below.

The ability to benefit from capital gains

A capital gain occurs when the value of an asset such as a house increases. Most consumer durable goods, such as cars and television sets, depreciate in value over the years after they are purchased. By contrast, the value of land and property generally increases, at least in the UK. Share prices also tend to rise in the long run, though they can also fall in value. On the whole, the already wealthy own the most expensive houses, often also owning more than one house, and they are the main owners of shares. The least wealthy often rent rather than own the houses they live in, and seldom own shares.

Private pension assets

Property and financial assets such as shares are generally forms of marketable wealth, accounting together for 48.59% of the household wealth in 2020,

shown in Figure 7.4. However, as Figure 7.4 shows, in 2020 private pension assets (the total value of all individuals' pensions not provided by the government), which are generally non-marketable, made up 42.34% of total household wealth. Until quite recently, company pension schemes were not available for many low-paid UK workers, and neither did low-paid workers contribute to private pension schemes sold, for example, by insurance companies such as Standard Life. This has meant that many low-paid workers have expected their retirement income to be provided solely by the state pension, which on its own is little more than a poverty income. (Application of Economics in the Real World 7.5 on page 221 explains the new automatic enrolment scheme which the UK government has introduced to extend private pensions to low-paid workers.)

Inheritance, gifts and luck

Wealthy families are often divided into 'new wealth' and 'old wealth'. Entrepreneurs who have built up large personal fortunes, often starting from scratch, through founding their own businesses and successful risk taking, fall into the category of 'new wealth'. Members of the landed aristocracy who pass wealth-holdings from generation to generation are 'old wealth'. 'Old wealth' can, of course, lead to 'new wealth' — for example, when the sons and daughters of rich families use their education and inherited wealth as the platform on which to develop entrepreneurial skills. Likewise, 'new wealth' can create 'old wealth', when for example the fortunes of the newly rich are passed on to the next generation.

Wealth taxation versus taxation of income

In the UK, a much larger fraction of the government's tax revenue comes from taxation of income than from the taxation of wealth. Wealth is lightly taxed and there are many loopholes through which the already wealthy can legally avoid paying wealth taxes such as inheritance tax. The wealthy can also afford to employ accountants and financial advisers who minimise the tax they are liable to pay, and in this way the wealthy become even wealthier.

The difference between equality and equity in relation to the distribution of income and wealth

KEY TERMS

equality when everyone is treated exactly the same. A completely equal distribution of income means that everybody has the same income.

equity when everyone is treated fairly.

The terms **equality** and **equity** are often used interchangeably, which can be confusing. Although the concepts are related, they are not the same. Complete equality in the distribution of income is achieved when each person receives exactly the same amount of income. The degree of inequality is indicated by the extent to which people's incomes differ. Equity is when people are treated fairly, but differently, having taken into account their different circumstances. Very few people would argue that it would be equitable if everyone received the same income, irrespective of their efforts and the contribution they make to society. It is when people start to discuss how much inequality is fair that the arguments usually begin.

Equity, which means fairness or justness, is a *normative* concept, which cannot be measured. Different people form different value judgements on what is equitable or inequitable. By contrast, equality is a *positive* concept. A positive statement, such as the view that the incomes received by everyone in the

SYNOPTIC LINK
We first described positive and normative statements in Chapter 1 on page 5.

economy are the same, can be tested against the evidence to see if it is true or untrue. As we will shortly explain, it is possible to measure the degrees of equality and inequality in the distributions of income and wealth.

TEST YOURSELF 7.3
Which of the following are normative statements?

A The distribution of wealth is more unequal than the distribution of income.

B The rich deserve to be rich.

C Inequalities can be measured.

D Governments ought to give more help to the poor.

EXTENSION MATERIAL

Horizontal and vertical equity

Government intervention in the economy, which treats people in the same circumstances equally, obeys the principle of *horizontal equity*. Horizontal equity occurs when households with the same income and personal circumstances (e.g. number of children) pay the same income tax and are eligible for the same welfare benefits. The related concept of *vertical equity* justifies taking income from the rich (on the grounds that they do not need it) and redistributing it to the poor (on the grounds that they do need it). The distribution of income after taxation and receipt of transfers is judged by many to be more equitable than the distribution of income before taxes have been paid and transfer payments received. (In this context, a *transfer* or *transfer payment* is a payment of money for which goods or services are not received in exchange. Governments use transfer payments, such as pensions, as means of income redistribution by giving out money as a part of their social welfare programmes.)

However, achieving greater vertical equity can conflict with another principle of intervention, the *benefit principle*, which argues that those who receive most benefit from government spending (e.g. motorists benefiting from roads) should pay the most in taxes.

STUDY TIP
Make sure you understand the difference between equity and equality.

The Lorenz curve and Gini coefficient

Economists use **Lorenz curves** and the **Gini coefficient** to measure inequality. A Lorenz curve measures *the extent* to which the distribution of income (or wealth) is equal or unequal. The degree of inequality is measured by a Gini coefficient statistic.

KEY TERMS
Lorenz curve a graph on which the cumulative percentage of total national income or wealth is plotted against the cumulative percentage of population (ranked in increasing size of share). The extent to which the curve dips below a straight 45-degree diagonal line indicates the degree of inequality of distribution.

Gini coefficient measures the extent to which the distribution of income or wealth among individuals or households within an economy deviates from a perfectly equal distribution.

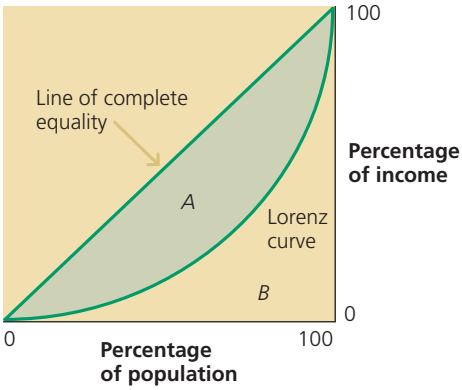


Figure 7.5 A Lorenz curve

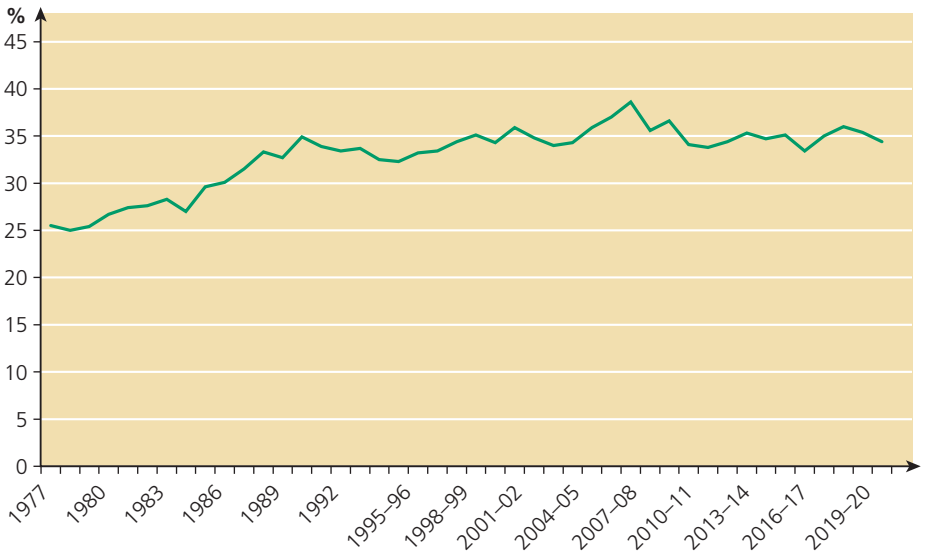
The Lorenz curve in Figure 7.5 shows population on the horizontal axis, measured in cumulative percentages from 0% to 100%. The vertical axis shows the cumulative percentage of income received by the population. If incomes were distributed equally, the Lorenz curve would lie along the diagonal line in the diagram. The nearer the Lorenz curve is to the diagonal, the more equal is the distribution of income.

The Gini coefficient measures the area between the Lorenz curve and the diagonal as a ratio of the total area under the diagonal. In terms of the diagram, the Gini coefficient is calculated using the following formula:

$$\text{Gini coefficient} = \frac{\text{area A}}{\text{area A} + \text{area B}}$$

The lower the value of the Gini coefficient, the more equally household income is distributed. If the Lorenz curve were to lie along the 45-degree line in Figure 7.5, every household would have exactly the same income and the Gini coefficient would be zero. At the other extreme, if one person received all the income and everybody else no income, the Lorenz curve would be the reverse L-shape lying along, first, the horizontal axis, and then the right-hand vertical axis in the diagram. Between these two extremes, Lorenz curves closer to the line of complete equality show greater equality (and Gini coefficients approaching zero), while Lorenz curves further away from the diagonal display greater inequality (and Gini coefficients approaching 1).

Figure 7.6 shows how the Gini coefficient for equivalised disposable income changed in the UK between 1977 and 2021. (Equivalised income is a measure of household income that takes account of the differences in a household's size and composition. It is used for the calculation of poverty.) Measured in this way, inequality of disposable income increased in the 1980s, and then fell in the early 1990s. After a slight increase around the millennium, it fell again from around 36% in 2001 to just over 34% in 2021.



Source: ONS

Figure 7.6 Gini coefficient for equivalised disposable income, UK, 1977–2021

TEST YOURSELF 7.4

Table 7.2 shows the distribution of income (by quintile) in country X and country Y. Plot Lorenz curves for the two countries.

Table 7.2 Income distributions by quintile

Quintile	Percentage of total disposable income	
	Country X	Country Y
1st	5	4
2nd	10	6
3rd	15	25
4th	20	30
5th	50	35
Total	100	100

TEST YOURSELF 7.5

Which two of the following are most likely to reflect a move towards a more equal distribution of income? Explain your answer.

A The share of total income going to the bottom 50% of income earners rises from 30% to 40%.

B The share of total income going to the top 15% of income earners rises from 60% to 62%.

C The Gini coefficient rises towards 1.

D The Gini coefficient falls towards 0.

The likely benefits and costs of more equal distributions of income and wealth

It is sometimes argued, particularly by economists who favour government intervention in markets, that more equal distributions of income and wealth can lead to faster economic growth. The fruits of faster growth can then be used to improve living standards and the economic welfare of most or all the population. The logic behind this argument is that people on low incomes and who possess little wealth generally spend all or most of their incomes on consumption and save very little, or nothing at all. Increased spending on consumer goods and services increases aggregate demand in the economy, which promotes economic growth. By contrast, the better-off spend a smaller fraction of their incomes on consumption, which leads to slower growth. Inequality can also mean that the talents of some people in society are wasted or, at least, not fully exploited. Inequality in income usually means that there is also inequality of opportunity.

However, economists of a more pro-free-market persuasion generally reject this view of the world, partly on the grounds that it takes no account of incentives and disincentives. They argue that the **progressive taxation** of higher incomes and wealth, combined with the transfer of taxed income to the less well-off, in the form of welfare benefits, significantly reduces the incentives to work hard, both among the better-off and among the poor. Reducing incentives to effort and enterprise reduces, rather than increases, the rate of economic growth. Extreme pro-free-market analysts argue that to make the poor better

KEY TERM

progressive taxation a tax is progressive when, as income rises, a greater proportion of income is paid in taxation. The term can be applied to a particular tax such as income tax or to taxation in general.

off *eventually*, inequality must first be increased rather than reduced. By incentivising those already in work to work harder, and the unwaged welfare benefit claimants to search for jobs, faster economic growth can be achieved. In the long run, the low-waged benefit even though in the short run their incomes fall.

Taken to the extreme, however, widening income and wealth inequalities undoubtedly increases poverty, at least in the short run. Most people in the UK seem to subscribe to a consensus view that there should be a certain amount of redistribution, compared to the outcome that would be achieved in a complete free-market situation. This redistribution is necessary for moral reasons and also to maintain social cohesion within society. The debating issues are, first, where to draw the line when redistributing income and wealth, and second, what methods should be used to achieve the degree of redistribution desired.

TEST YOURSELF 7.6

If a country's Gini coefficient fell from 0.9 to 0.7 over a five-year period, what would this tell you about the distribution of income within the country?

STUDY TIP

Make sure you understand, first, the difference between income and wealth, and second, the difference between equal and equitable.

SYNOPTIC LINK

A detailed explanation of the main types of tax is given in Chapter 13 on fiscal policy, especially pages 479–80. This will help you to understand the main differences between proportional, progressive and regressive taxation.

STUDY TIP

Make sure you understand the links between equality and inequality on the one hand and poverty on the other. There is significant overlap between the causes and effects of inequality and poverty which you will need to appreciate when answering questions on these topics.

QUANTITATIVE SKILLS 7.3

Worked example: calculating the mean and the median

The 12 fellow students in your economics class have told you that in a typical week their household incomes are £480; £750; £300; £250; £1,000; £600; £400; £500; £800; £520; £550; and £450. Your own household income is £570. What is the mean household income and the median household income in your class?

The mean and the median are two kinds of 'average', with the mean being the most commonly recognised. The mean household income can be calculated by adding up the household incomes (in this case £7,170) and dividing this number by 13 (the total number of households in this example, as your household is included as well as those of your fellow students). The mean value for all 13 households is approximately £551.54.

The 'median' is the 'middle' value in the list of numbers. To find the median, you should list the numbers in numerical order (£250; £300; £400; £450; £480; £500; £520; £550; £570; £600; £750; £800; and £1,000), and then locate the middle value. In this example, the median value is £520.

APPLICATION OF ECONOMICS IN THE REAL WORLD 7.2

The case for and against a flat tax

A 'flat tax' is the same as a proportional tax. Pro-free-market economists in the USA and elsewhere have been arguing the case for a 'flat tax' for decades. In 1994, Estonia became the first country in Europe to introduce a flat tax, replacing the existing three tax rates on personal income, and another on corporate profits, with one uniform rate of 26%. Other countries in central and eastern Europe have since followed suit, some introducing flat tax rates as low as 12 or 13%. In the UK, the Conservatives have considered but not taken up the idea.

The main case for flat taxes is that fewer brackets are simpler to administer, with one tax bracket being the simplest of all. If a pure flat income tax were to be introduced in the UK, HM Revenue and Customs would take the same amount from the first pound earned by a taxpayer as would be taken from the last pound earned. As a result, many high-income earners want flat taxes to be introduced.



Tallinn, the capital of Estonia, where a flat tax was introduced in 1994

Rory Meakin, who is the research director at the pro-free-market lobby group, the Taxpayers' Alliance, is in favour of flatter taxes and a much simpler tax system. Meakin argues:

> Britain's tax code is one of the longest in the world. We have a basic rate, a higher rate and an additional rate of income tax [at that time], with a different set of rates for dividends and yet another for savings.
>
> There's no need for taxes to be so maddeningly complex. Complexity also makes the tax system both economically and socially destructive. Quite apart from the money it sucks out of the productive economy, it's economically damaging for two reasons. First, it requires an army of clever accountants and tax lawyers to navigate the system for businesses and an opposing battalion of bureaucrats to monitor all those rules and rates. Second, it means that the public just doesn't understand how the tax system operates, which, among other things, risks discouraging people with new commercial ideas from starting new businesses. That means opportunities are lost and fewer jobs are created. And it's socially damaging because the confusion leads to suspicion that others are getting away with something, directly leading to distrust of the whole system.
>
> We need to sweep away all that complexity and replace the current system of taxation with lower, simpler and more proportionate taxes which are easier to understand, with fewer loopholes and exemptions.

By contrast, Richard Murphy, head of the anti-poverty Tax Research partnership, which is funded by the anti-poverty Joseph Rowntree Charitable Trust, is against flat taxes. Murphy believes that the view that flat taxes are simple and would raise tax revenue is a myth, as is the view that a great deal of the cost of administering the tax system could be eliminated. This would only be true if taxes were scrapped along with the money they raise. This provides the real clue to what flat taxes are all about. It is not chance that flat taxes are always promoted by people who also argue for small government and massive cuts in public spending.

Follow-up questions

1. Find out the difference between a flat tax or proportional tax, a progressive tax and a regressive tax.
2. Giving an example of each, explain the difference between tax avoidance and tax evasion.
3. Supporters of flat taxation believe that flat taxes would reduce tax avoidance and tax evasion. Explain why you agree or disagree with this view.
4. Evaluate the case for massive cuts in public spending in the UK.

SECTION 7.1 SUMMARY

- The distribution of income measures how personal or household income is distributed among different income groups in society, such as between rich and poor.
- The distribution of wealth measures how personal or household wealth is distributed among different groups in society.
- Inequalities in the distribution of UK income and wealth have tended to widen over recent decades.
- Inequalities in the distribution of wealth have always been wider than those in the distribution of income.
- An important factor influencing the distribution of income is the shares of national income going to different factors of production.
- Other factors affecting the distribution of income include the shares of earned and unearned income, the widening of earnings differences between workers at the top and bottom of the pay league table, and the effect of globalisation and the international migration of workers.
- Factors influencing the distribution of wealth include the ability to benefit from capital gains, ownership of private pension assets, the roles of inheritance, gifts and luck, and the different effects of wealth and income taxation.
- Equality must not be confused with equity or fairness.
- Economists use Lorenz curves and the Gini coefficient to measure inequality.
- A Lorenz curve measures *the extent* to which the distribution of income (or wealth) is equal or unequal.
- The degree of inequality is measured by a Gini coefficient statistic.
- Economists who favour government intervention in markets believe that more equal distributions of income and wealth can lead to faster economic growth. The fruits of faster growth can then be used to improve living standards and the economic welfare of most or all the population.
- Economists of a pro-free-market persuasion believe that government intervention reduces the incentives to work hard, both among the better-off and among the poor.

7.2 The problem of poverty

KEY TERM
poverty the state of being extremely poor and not having enough money or income to meet basic needs.

Poverty is the state of being extremely poor and not having enough money or income to meet basic needs including food, clothing and shelter. The World Bank describes poverty in the following way:

> Poverty is hunger. Poverty is lack of shelter. Poverty is being sick and not being able to see a doctor. Poverty is not having access to school and not knowing how to read. Poverty is not having a job, is fear for the future, living one day at a time.

Poverty is caused both by a low real national income relative to a country's total population size and by inequalities in the distributions of income and wealth. The former leads to *absolute* poverty for many if not most of a country's inhabitants, whereas the latter causes *relative* poverty. These terms are explained below.

The difference between relative and absolute poverty

Absolute poverty

KEY TERM
absolute poverty a condition characterised by severe deprivation of basic human needs, including food, safe drinking water, sanitation facilities, health, shelter, education and information. It depends not only on income but also on access to services.

The Joseph Rowntree Foundation defines **absolute poverty** in the following way:

> A poverty level that does not change over time, in terms of the living standard that it refers to. It stays the same even if society is becoming more prosperous. An absolute poverty line thus represents a certain basic level of goods and services, and only rises with inflation to show how much it would cost to buy those goods and services.

Relative poverty

KEY TERM
relative poverty occurs when income is below a specified proportion of average income, e.g. below 60% of median income.

Relative poverty is suffered by a household if its income is below a specified proportion of the average income for all households. For example, in the UK, the 'poverty line' has usually been set at 60% of median income. The charity Barnardo's says:

> Relative poverty defines 'poverty' as being below a relative poverty threshold. It classifies individuals or families as 'poor' not by comparing them to a fixed cut-off point, but by comparing them to others in the population under study.

TEST YOURSELF 7.7
Explain why statement C provides the correct answer and why the other possible options are wrong.

The inability to satisfy basic food needs is part of the definition of:

A relative poverty

B obesity

C absolute poverty

D idleness

APPLICATION OF ECONOMICS IN THE REAL WORLD 7.3

How the Conservatives tried to redefine child poverty

On 1 July 2015, Iain Duncan Smith, the then secretary of state for work and pensions, announced that he intended to repeal the 2010 Child Poverty Act, which committed the government to a target of eradicating child poverty in the UK by 2020. The government wanted to get rid of the definition of relative poverty in the 2010 Act (anyone in a household beneath 60% of median income), to abandon all the 2010 targets, and to introduce a new definition of relative poverty, embracing work and education levels in the family.

The work and pensions secretary wanted to remove a statutory duty to publish levels of UK household income as part of the welfare reform. However, following fierce opposition led by the Child Poverty Action Group (CPAG) and the House of Lords, the government did a 'U-turn' and accepted that the material deprivation measures should remain protected.

Follow-up questions

1. Explain the difference between relative poverty and absolute poverty.
2. Access the CPAG's website and find out more detail about what has happened to Duncan Smith's proposals since 2015.
3. Explain why you agree or disagree with the proposed new definition of child poverty which has now been abandoned.
4. The growth in the number of people sleeping on the streets is a symptom of growing poverty in the UK. Between 2010 and 2021, rough-sleeping estimates showed an increase of 38% (source: www.gov.uk). Evaluate one way in which rough-sleeping could be reduced.

The causes and effects of poverty in the UK

For the most part, UK poverty is relative rather than absolute, although many homeless people living in cardboard boxes in town centres are absolutely as well as relatively poor. Three of the main causes of relative poverty in the UK are old age, unemployment and the low wages of many of those in work.

Old age and poverty

Old age causes relative poverty largely because, as we have seen, many people of state pension age rely on the state pension and lack a private pension. Before

SYNOPTIC LINK
Some of the key price indices, such as the retail prices index (RPI) and consumer prices index (CPI), are covered in more detail in Chapter 9.

the early 1980s, the state pension rose each year in line with the increase in national prosperity delivered by economic growth and higher real earnings. However, pensions then moved away from being linked to average earnings, to be linked instead to the rate of inflation, measured at first by changes in the retail prices index (RPI). This method of calculating pensions is known as 'index-linking', where they rise at the same rate as the average price level in an economy. Pensions no longer rose in line with the general increase in the standard of living and pensioners' living standards were in effect kept at their 1980 level. Older people who were reliant solely on the state for their income became increasingly worse off compared to people in work.

In 2011, the Coalition government introduced what is known as the 'triple lock'. The government said that the state pension would now rise in line with either earnings, or CPI (consumer prices index) inflation, or a 2.5% increase, whichever is the greatest. Shortly afterwards, the government considered switching back to linking pension increases solely to rises in average earnings, thus ending the annual inflation-linked rise in benefits. In contrast to earlier decades, when the rate of increase in average earnings exceeded the rate of inflation, by 2013 earnings were rising by around 1% a year, which was lower than the rate of inflation — which, at the time, was nearer to 2%. By 2015, average earnings were rising by slightly more than the then low rate of inflation. Following a dip around the time of the Brexit vote in 2016, by 2018 average earnings had resumed their slight increase. However, whatever is likely to happen in the future, unless supplemented by other sources of income, the state pension will continue to be very much a poverty income.

Unemployment and poverty

Unemployment benefits are generally lower than the pay workers received before losing their jobs. An increase in unemployment therefore increases poverty. In 2013, as part of the policy of cutting public spending, the Coalition government introduced a limit on the total amount of benefit that most people aged 16 to 64 can claim. This limit may further increase poverty, though this effect could be offset by a continuing rapid fall in unemployment. Absolute poverty can best be reduced by fast and sustained economic growth and by creating jobs. This may also reduce relative poverty. Economic growth can also create the additional financial resources for governments to increase the real value of the state pension and unemployment benefits. Possibly in part due to the political power of the so-called 'grey vote', the government may be more minded to increase the real value of the state pension than of unemployment benefits.

Low wages and poverty

When discussing the nature of both absolute and relative poverty in the UK, it is important to distinguish between the low-waged and the unwaged, who are unemployed. The low-waged, unlike the unwaged, are workers with jobs, albeit jobs in which their hourly and weekly earnings are low. The low-waged include many unskilled workers, together with skilled workers who have lost their jobs in industrial sectors such as manufacturing and coal mining, and who have had to accept low-waged, unskilled employment. The low-waged poor are almost always relatively poor rather than absolutely poor. In contrast, some of the unwaged, including homeless people living on the street, fall into the category of the absolutely poor. The national minimum wage, which we explained in Chapter 6, was an attempt to reduce the poverty of the low-waged. The national living wage has now taken on this role.

TEST YOURSELF 7.8
How does the education that people receive affect the chances of whether they are rich or poor?

The effects of poverty in the UK

According to the Child Poverty Action Group (CPAG), for some people, poverty damages childhoods, life chances and eventually all in society. It is a direct cost to government resulting from additional demands placed on services and benefits for the poor, as well as reduced tax receipts.

Educational deprivation

In terms of educational deprivation, the CPAG argues that children from poorer backgrounds lag behind at all stages of education. For example, by the age of 3, poorer children are estimated to be, on average, nine months behind children from wealthier backgrounds. By the end of primary school, pupils receiving free school meals (an indicator of those children who are from poorer backgrounds) are estimated to be almost three terms behind their more affluent peers. This gap grows to over five terms by the age of 14, and by 16, children receiving free school meals achieve 1.7 grades lower at GCSE.

Health deprivation

The CPAG says that poverty is also associated with a higher risk of both illness and premature death. Children born in the poorest areas of the UK weigh, on average, 200 grams less at birth than those born in the richest areas. Children from low-income families are also more likely to die at birth or in infancy than children born into richer families. They are more likely to suffer chronic illness during childhood or to have a disability. Poorer health over the course of a lifetime has an impact on life expectancy: professionals live, on average, eight years longer than unskilled workers.

The effect of poverty on communities

Children living in poverty are almost twice as likely to live in bad housing. This has significant effects on both their physical and mental health, as well as their educational achievement. Fuel poverty also affects children detrimentally as they grow up. Low-income families sometimes have to make a choice between food and heating. Children from low-income families often forgo activities that most children would take for granted. They miss school trips; can't invite friends round for tea; and can't afford a one-week holiday away from home.



Poverty is often associated with poor-quality housing and adverse effects on people's health

Besides having serious effects on children, poverty adversely affects other groups in society, particularly pensioners whose sole income is the state pension. For example, the elderly are the main group in society suffering from fuel poverty. Though in 2016 energy prices fell for a time, falling energy prices generally turn out to be a short-term 'blip', soon to be replaced by the 'normal service' of rising prices in response to global resource depletion. During 2022 energy companies announced price hikes that pushed up prices right across the board. Old-age poverty and the fact that people are now living for longer lead to other adverse effects on society — for example, elderly 'bed-blockers' continuing to occupy hospital wards after they have been successfully treated, simply because they have nowhere else to go.

SECTION 7.2 SUMMARY

- Poverty is caused both by a low real national income relative to a country's total population size and by inequalities in the distributions of income and wealth.
- The former leads to absolute poverty for many if not most of a country's inhabitants, whereas the latter causes relative poverty.
- Absolute poverty is a condition characterised by severe deprivation of basic human needs, including food, safe drinking water, sanitation facilities, health, shelter, education and information. It depends not only on income but also on access to services.
- Relative poverty occurs when income is below a specified proportion of average income: for example, below 60% of median income.
- Relative poverty is more significant than absolute poverty for most of the poor in the UK, except the destitute.
- Old age, unemployment and low wages are important causes of poverty in the UK.
- The Child Poverty Action Group believes that poverty damages childhoods, life chances and eventually all in society. It leads to health and education deprivation and to the breakdown of communities.
- The old form a significant part of the poor.

7.3 Government policies to alleviate poverty and to influence the distribution of income and wealth

Poverty and the tax and benefits system

Poverty is seldom caused directly and immediately by taxation and the benefits system. However, through a process known as *fiscal drag* (which will be explained shortly) and through cuts in welfare benefits, poverty can increase. Making taxation more progressive and increasing welfare benefits reduces poverty and inequalities in the distribution of income, at least in the short run. However, as we have already mentioned, pro-free-market economists believe that, in the drive to reduce inequality, these changes worsen labour market

incentives, competitiveness and economic growth. They may lead to a culture of welfare dependency where some people are long-term unemployed and are reluctant to look for work. According to this view, in the long run, low incomes may fail to grow, and poverty may increase. If true, government intervention in labour markets in an attempt to reduce poverty results in government failure.

STUDY TIP

Make sure you understand the concepts of market failure and government failure, which are explained in Chapter 8. Inequality and poverty may be seen as significant forms of market failure, while government methods to tackle them may make the problems worse, or lead to new problems. These are important issues for you to evaluate in an exam question.

EXTENSION MATERIAL

Universal benefits versus means-tested benefits

Government welfare benefits paid to the poor and to other groups in society can be divided into: national insurance benefits, such as state pensions; universal benefits; and means-tested benefits. The winter fuel payment (financial assistance from the government to pay for heating bills), which in 2022 ranged between £100 and £300 (plus additional payments to address the fuel cost crisis), depending on personal circumstances, is a universal benefit: it is paid to all pensioner households who claim the benefit, irrespective of household income. Winter fuel payment is also untaxed. (Pensioners 'opt in' to claim winter fuel payment. However, some behavioural economists argue that rich pensioners who currently claim the payment should be 'nudged' or encouraged to opt out of continuing to claim the benefit, on the grounds that they do not need it.)

Over the years, UK governments have sought to replace universal benefits with means-tested benefits. This is because, as is the case with the winter fuel payment, benefits that are not means-tested are usually expensive to provide and go to people who do not need them. (The growth of means-tested benefits also reflects the fact that other benefits, including national insurance benefits, have not been increased in line with earnings.) However, universality means that people who do need the benefits usually end up receiving them. With means testing, by contrast, people who need a benefit, particularly the elderly poor, may be too proud to claim the benefit, or be put off by the amount of form filling needed to prove eligibility.

With means-tested benefits, people's ability to claim a benefit is determined by their income and their outgoings (or 'means'). Means-tested benefits are more efficient than universal benefits in the sense that taxpayers' money is less likely to be given to people who do not need the benefit. However, as we will shortly see in the context of the poverty and unemployment traps, means testing can create unintended disincentive effects which prevent the low-waged from working harder and the unwaged (or unemployed) from seeking employment.

Between 2015 and the time of writing (in 2022), the Conservative government has been 'rolling out' Universal Credit (UC) to replace a number other means-tested benefits, including Housing Benefit, Income Support and income-related Jobseeker's Allowance, which is the main benefit paid to the unemployed. Originally announced in 2013, the introduction of Universal Credit has been hindered by very expensive ICT problems, and it will be several years before it has fully replaced the separate benefits we have just listed.

TEST YOURSELF 7.9

Is Universal Credit a universal benefit or a means-tested benefit?

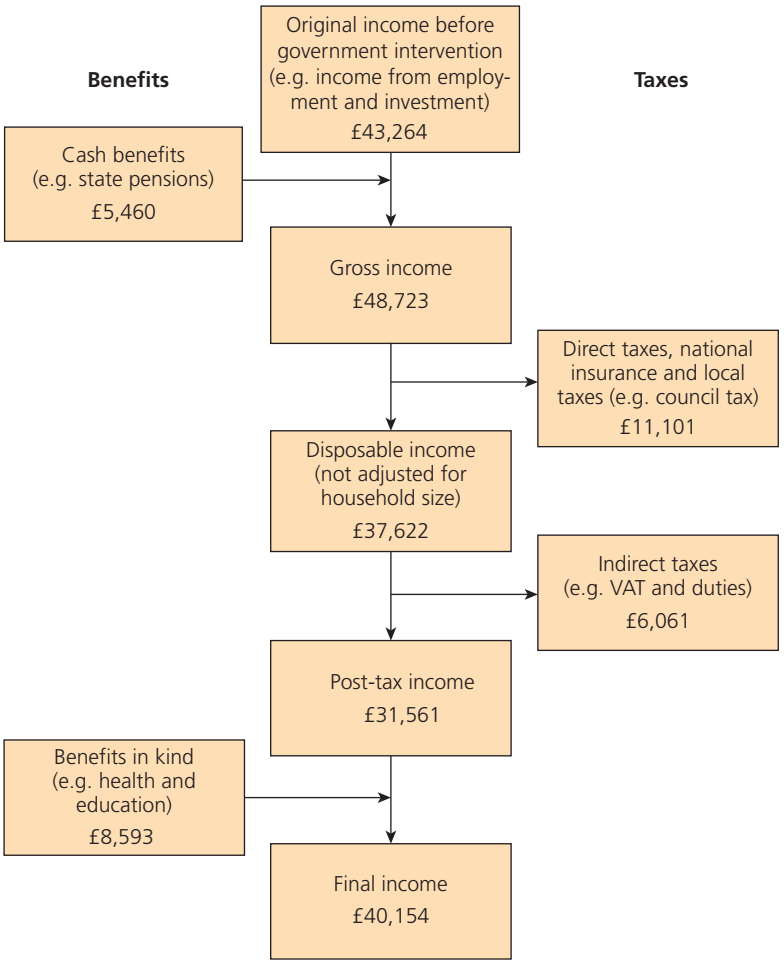
The effects of taxes and benefits on household income, 2020/21

A household is a group of people, such as a family, who are living together in a house or flat. According to the Office for National Statistics, in 2019/20 before direct taxes were paid and cash benefits received, the richest fifth had an average original income 12 times larger than the poorest fifth — £131,670 per year compared with £10,958. When direct taxes and cash benefits were

included, income was shared less unequally between households — the income of the poorest fifth of households increased by £5,677 to £16,635, and the income of the richest fifth decreased by £37,102 to £94,568.

Consequently, the ratio of income of the richest fifth to the poorest fifth fell from 12 to 1, to just under 6 to 1. The inclusion of indirect taxes (such as alcohol duties and value added tax) and benefits-in-kind (e.g. state education and the National Health Service) further reduces this ratio to less than 4 to 1 (source: ons.gov.uk).

This fall in the ratio of income between the richest fifth and the poorest fifth of households is explained by comparing the composition of taxes and benefits of these two groups of households. The poorest fifth of households received relatively larger amounts of both cash benefits and benefits-in-kind in 2019/20. Richer households, on the other hand, paid higher amounts in taxes — both direct and indirect.



Source: ONS

Figure 7.7 Average household income, cash benefits received, including benefits in kind, and taxes paid, 2020/21

There are five stages, illustrated in Figure 7.7 for the tax year 2019/20, through which taxes and benefits affect the distribution of income. These are as follows:

1 Household members begin with income from employment, private pensions, investments such as the shares they own, and other non-government sources. This is referred to as 'original income'.
2 Households then receive income from cash benefits. The sum of cash benefits and original income is referred to as 'gross income'.

3 Households then pay direct taxes. Income after direct taxes have been subtracted from gross income is called 'disposable income'.
4 Indirect taxes are then paid on spending on goods and services. Disposable income minus indirect taxes is referred to as 'post-tax income'.
5 Households finally receive a benefit from services provided by the government (benefits in kind). Post-tax income plus benefits in kind is referred to as 'final income'.

SYNOPTIC LINK
Government policies to deal with poverty are examples of fiscal policy. See Chapter 13 for detailed discussion of the UK government's fiscal policy.

TEST YOURSELF 7.10
What is meant by the terms 'original income' and 'disposable income'?

QUANTITATIVE SKILLS 7.4

Worked example: plotting data on a bar chart

Table 7.3 shows workers in poverty by employment type for the financial years from 2009/10 to 2016/17.

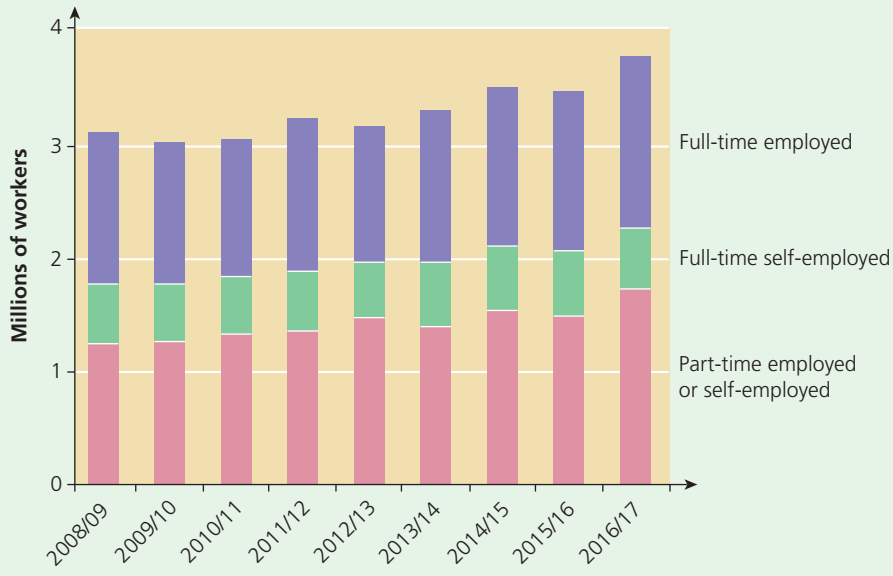
Table 7.3 Workers in poverty by employment type (millions)

Year	Full-time employed	Full-time self-employed	Part-time employed or self-employed
2009/10	1.26	0.52	1.27
2010/11	1.28	0.53	1.32
2011/12	1.36	0.52	1.36
2012/13	1.21	0.50	1.47
2013/14	1.36	0.59	1.39
2014/15	1.39	0.57	1.54
2015/16	1.42	0.57	1.49
2016/17	1.51	0.55	1.72

Plot the data on a bar chart.

The answer is shown in Figure 7.8.

There are various ways of using bar graphs or charts. For answering the question, we have used a stacked bar chart. Stacked bar charts are a way of showing information about different sub-groups of the main categories.



Source: ONS

Figure 7.8 Workers in poverty by employment type

Fiscal drag, poverty and low pay

The UK tax system has affected poverty partly through a process known as fiscal drag. Fiscal drag occurs in a progressive income tax system when the government fails to raise tax thresholds (or personal tax allowances) to keep pace with inflation. Figure 7.9(a) shows an income pyramid with the rich at the top and the poor at the bottom, and with the tax threshold fixed at an income of £10,000. In this example, a person with an income of £9,900 is just below the threshold and pays no income tax.

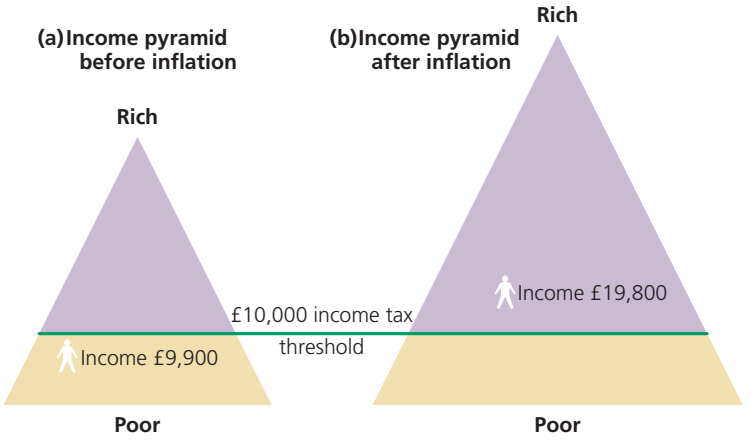


Figure 7.9 How fiscal drag causes real income to fall as tax thresholds do not rise in line with inflation

Suppose that over the next few years both prices and all money incomes exactly double. In the absence of taxation and assuming there are no benefits available, real incomes would remain unchanged, with households no better or worse off. But if the government fails to increase personal tax allowances in line with inflation (i.e. to raise the tax threshold to £20,000), a doubling of the person's money income to £19,800 means that £9,800 of income is now taxable. The individual concerned is now worse off in real terms.

The new situation is shown in Figure 7.9(b). Inflation has dragged the low-paid worker across the basic tax threshold and they must now pay income tax. In a similar way, higher-paid workers have to pay a higher proportion of their income in tax if the higher 40% and 45% **marginal tax rates** remain unadjusted for inflation.

In recent years in the UK, the Conservative government has tried to reduce relative poverty among the low-paid by raising income tax thresholds by more than the rate of inflation, taking the basic tax threshold to £12,570 in 2022. The increases in the personal tax allowance, or tax threshold, that have been made have taken a significant number of the low-paid out of the income tax net and 'clawed back' some of the fiscal drag that took place in earlier years.

KEY TERM
marginal tax rate the tax rate levied on the last pound of income received. The term can be applied solely to income taxes or to all the taxes that a person or business pays.

TEST YOURSELF 7.11
What is the difference between a tax threshold and a tax ceiling?

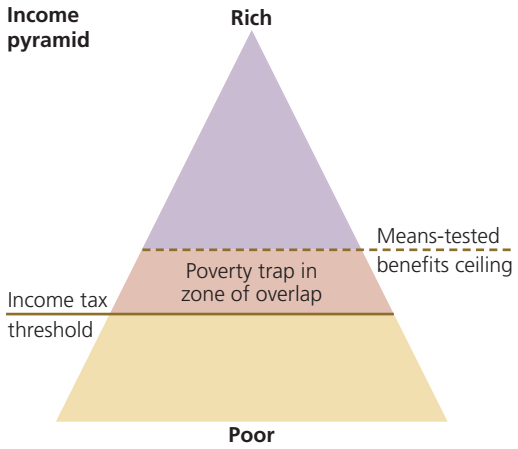


Figure 7.10 The earnings trap or poverty trap

The earnings trap or poverty trap

Fiscal drag is one of the causes of the poverty trap. As there are a number of ways in which the poor can be trapped in poverty, this particular trap, which traps the *low-waged* in relative poverty, is perhaps better called 'the earnings trap'. It affects people in employment on low rates of pay, rather than the unemployed who are *unwaged*. Another cause of poverty affecting the homeless stems from the fact that to get a job they need a home, but to get a home they first need a job.

The immediate cause of the earnings trap is the overlap, illustrated in Figure 7.10, between the income tax threshold (the level of income at which income tax starts to be paid) and the means-tested welfare benefits ceiling (the level of income at which means-tested benefits cease to be paid). When welfare benefits are means-tested, a person's right to claim the benefit is reduced and eventually tapers away or disappears completely, as income rises. By contrast, a universal benefit is claimed as of right and is not dependent on income, although the state may decide to 'claw back' universal benefits that are taxable.

Low-paid workers caught within the zone of overlap in Figure 7.10 not only pay income tax and national insurance contributions on each extra pound earned; they also lose part or all of the right to claim benefits. Thus low-paid workers and their families whose income falls within this zone of overlap become trapped in relative poverty, since any increase in their pay results in little or no increase (and in extreme cases a fall) in their disposable income.

The effective marginal rate of taxation of workers in poorly paid occupations is high when the loss of means-tested benefits is added to deductions through income tax and national insurance contributions. Calculated in this way, the marginal tax rates of the low-paid can be much higher (often around 70% and in extreme cases over 100%) than the top 45% rate currently paid by the better-off on income (in 2022) above £150,000. Moreover, since the low-paid are generally employed in occupations yielding little job satisfaction or scope for legal tax avoidance, it can be argued that disincentives to work imposed by the UK tax and benefits system affect the poor at the lower end of the income pyramid much more than they affect the better-off.

The poverty trap can be eliminated by getting rid of the zone of overlap in the income pyramid illustrated in Figure 7.10. The income tax threshold could be raised to take low-waged households out of the tax net. Means-tested benefits could be completely replaced by universal benefits, though this is unlikely to happen. The national living wage rate might also reduce poverty by being raised to a higher rate than is currently planned in an attempt to prevent employers paying 'poverty wages'. However, raising the NLW could be counterproductive if unemployment increases as a result. At the time of its introduction, the Conservative government said that the NLW might lead to a loss of 60,000 jobs, which implies that a higher NLW rate would lead to even more job losses. However, the government also said that continuing economic growth would create many more jobs.

APPLICATION OF ECONOMICS IN THE REAL WORLD 7.4

Universal Credit has probably placed more families in the poverty trap

When Universal Credit was announced in 2013, the UK government stood accused that its changes to the tax and benefits system risked penalising almost 2 million low earners. A report by the anti-poverty campaigning group, the Joseph Rowntree Foundation, suggested that with many left to deal with a more complex benefits system than before, Universal Credit could see people worse off in work and struggling to manage their finances. The report stated that, while 'making work pay' is the key aim of Universal Credit, many households are set to be worse off, or only marginally better off.

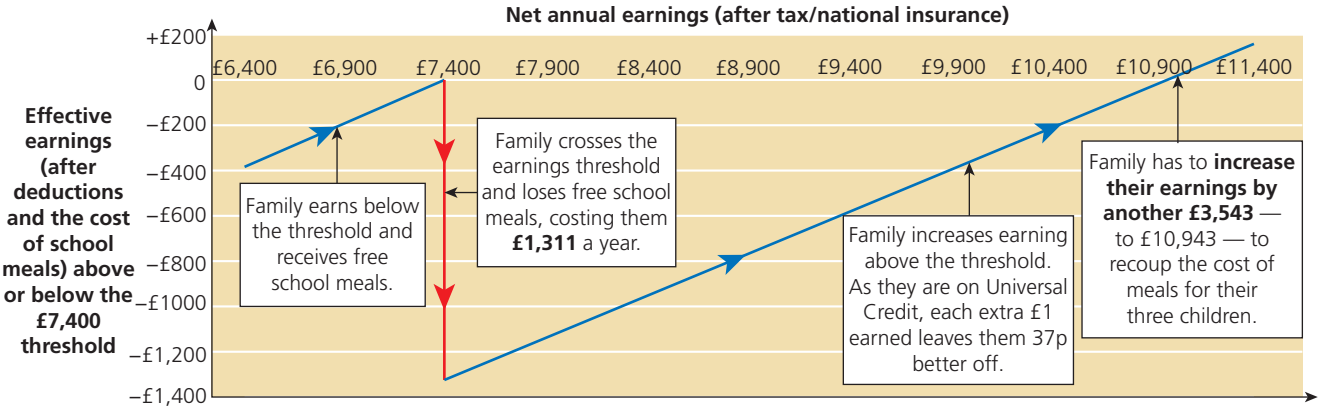
The report also raised serious concerns about a 'one-size-fits-all' IT-based delivery system and about potential IT failures which could quickly lead to backlogs, poor service and complaints.

However, the government claimed that in monetary terms, 3.1 million households would be entitled to more benefits as a result of the introduction of Universal Credit, while 2.8 million households would be entitled to less. Iain Duncan Smith, the secretary of state for work and pensions who introduced Universal Credit, said that across all households, there would be an average household gain of £16 per month.

One of the principal goals of Universal Credit is to ensure that additional earnings will always leave families better off and to avoid 'cliff edges' in incomes. In the foreword to the Universal Credit White Paper (a government report), Iain Duncan Smith said:

> At its heart, Universal Credit is very simple and will ensure that work always pays and is seen to pay. Universal Credit will mean that people will be consistently and transparently better off for each hour they work and every pound they earn.

But in 2018 the introduction of a £7,400 earnings ceiling, which limited a family's right to claim free school meals for their children, created a serious cliff edge that actually increased family poverty. According to the Children's Society, the effect of the cliff edge is illustrated in Figure 7.11.



Source: Children's Society and Child Poverty Action Group

Figure 7.11 The free school meals 'cliff edge' and poverty trap for a family with three children

Follow-up questions

1 Explain how Figure 7.11 illustrates the 'cliff edge' described in the text.
2 Find out whether, at the time you read this box, the 'roll out' of Universal Credit has been completed.
3 Compare a universal benefit with a means-tested benefit.
4 Discuss whether Universal Credit has reduced the poverty trap, or increased poverty and drawn more families into the trap. To help your answer, access the web page **https://fullfact.org/education/free-school-meals-and-universal-credit/** which investigates the claim by the Children's Society quoted above and illustrated in Figure 7.11.

The unemployment trap

The poverty trap, described earlier, affects the low-waged in jobs rather than the unemployed who are unwaged. It is important not to confuse the poverty trap or earnings trap with the unemployment trap. The unemployment trap

STUDY TIP

Make sure you understand the difference, and also the relationship, between the poverty trap (or earnings trap) and the unemployment trap.

is closely related to the earnings trap, since both affect the poor and result from the nature of the tax and benefits systems. But people caught in the unemployment trap are out of work — at least in terms of officially declared employment. The unemployment trap contains unwaged social security claimants who choose unemployment. This is because they decide they are better off out of work, living on benefits, than in low-paid jobs paying income tax and national insurance contributions (NICs), and losing some or all of their right to claim means-tested benefits.

One link between the earnings trap and the unemployment trap is the underground economy — the hidden or informal economy in which people work, usually for cash payments, while failing to declare income and sometimes fraudulently claiming social security benefits. Low-paid workers in employment can escape the earnings trap by giving up declared work in order to claim unemployment benefits, while receiving income from undeclared work in the underground economy. The underground economy is sometimes called the 'black economy'.

APPLICATION OF ECONOMICS IN THE REAL WORLD 7.5

Enrolling low-paid workers into workplace pension schemes

As they approach retirement, some workers wrongly believe that their entitlement to a state pension will provide them with more than just a 'poverty income'. As a result, until recently too many have chosen not to opt into private pension plans. Two further factors have contributed to under-enrolment. In the first place, many low-paid and often young workers saw that contributing to a private pension scheme would leave them with even less of their already meagre incomes to spend on everyday living. Better to adopt a 'live now, pay later' approach to life and to spend now and not bother to save.

In the second place, private pension schemes were until recently often based on the 'opt-in' principle. Lethargy and inertia meant that too few workers bothered to join a private pension scheme.

Drawing on the insights of behavioural theory which we explained in Chapter 2, this is now changing. The UK government has introduced, since 2012, a system of automatic workplace pension contributions, in which a slice of workers' pay packets is diverted to savings pots to pay for their eventual pensions. Employers are obliged to pay in as well, and the government also adds a little extra through tax relief.

The new system, called automatic enrolment, is an important example of 'nudge' theory being taken on board in government policy. Unless employers already provide 'in-house' workplace pension schemes (which have traditionally been available mostly for the better-paid professional workers), workers are now being automatically enrolled in workplace pension schemes — unless they choose to 'opt out'.

The 'opt-out' scheme was gradually introduced between 2012 and 2018. It is expected that relatively few eligible workers will decide to opt out. Once again, this is partly due to lethargy, but workers also know opting out will lead to the loss of employers' contributions. Workers below the age of 22 will still have to opt in, as will part-time workers. Since the pension contributions of low-paid workers are generally small, the value of their pensions was expected to be £6,549.40 a year in 2018/19 compared to £6,360 a year in 2017/18. When the private pensions have been added to the state pension, the retirement incomes of previously low-paid workers will still be relatively low compared to more fortunate pensioners. Automatic enrolment may also have the unintended consequence of hastening the introduction by so-called rogue employers of more casual part-time employment and zero-hours employment contracts.

Follow-up questions

1. Explain the difference between 'opting in' and 'opting out' in the context of a pension scheme.
2. What is a zero-hours employment contract?
3. Distinguish between the state pension and a private pension.
4. Why is pension automatic enrolment an example of the behavioural economics concept of 'nudge'?

SYNOPTIC LINK

Application of Economics in the Real World 7.5 illustrates how the insights of behavioural economics covered in Chapter 2 can be used to evaluate the effects of government economic policy.

The consequences of government policies which affect poverty and the distribution of income and wealth

We have made passing reference to some of these consequences in earlier sections of this chapter. Among the points we have made are:

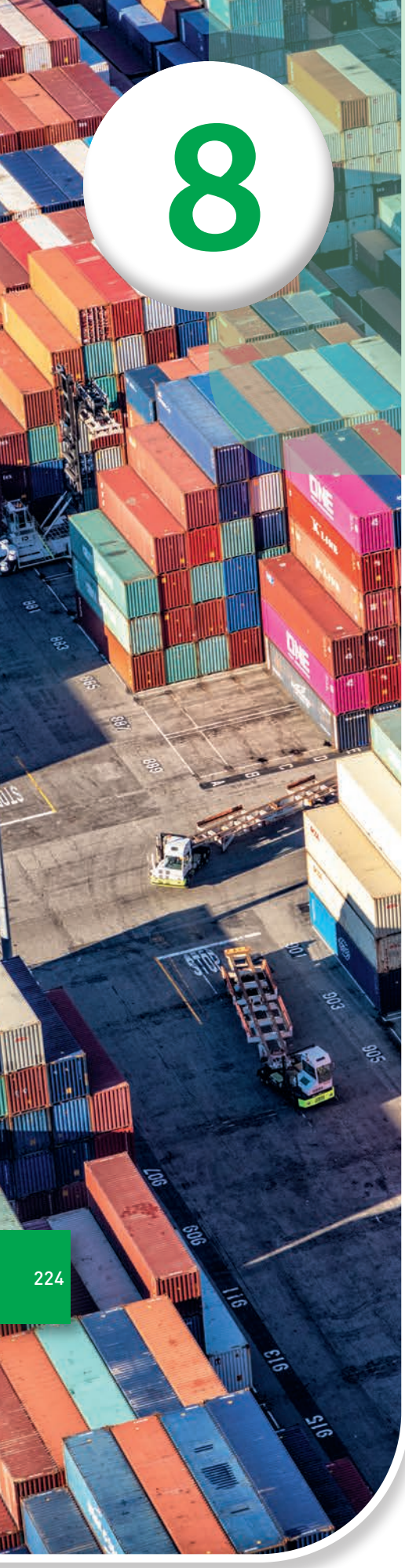
- Redistributive policies can make the distributions of income and wealth more equal and reduce relative poverty. Many would argue that such policies are equitable. However, by reducing incentives to work hard and to be entrepreneurial, redistributive policies may have adverse unintended consequences, such as slower economic growth and loss of international competitiveness.
- Redistributive policies can alleviate child poverty, old-age poverty and fuel poverty, but these aims have been only partially achieved.
- Arguably, faster economic growth provides the main avenue for reducing absolute poverty, and perhaps also relative poverty, but according to pro-free-market economists, wider inequalities in the distributions of income and wealth may be necessary if fast growth is to be achieved and sustained.
- Some fiscal measures, used to try to reduce poverty and narrow inequalities in the distribution of income, have resulted in the development of the poverty and unemployment traps.

SECTION 7.3 SUMMARY

- In the past, governments have used progressive taxation and transfers to the poor as the main methods of making the distribution of post-tax income more equal.
- Fiscal drag occurs in a progressive income tax system when the government fails to raise tax thresholds (or personal tax allowances) to keep pace with inflation.
- Fiscal drag contributes to the existence of a poverty trap or earnings trap, and also to an unemployment trap.
- It is important to understand the difference, and also the relationship, between the poverty trap (or earnings trap) and the unemployment trap.
- The poverty trap affects people in employment on *low rates of pay*, rather than the unemployed who are *unwaged*.
- The unemployment trap contains unwaged social security claimants who choose unemployment.
- Government intervention to reduce poverty may lead to government failures that make matters worse.
- Pro-free-market economists focus on the disincentive effects of progressive taxation and transfers, whereas interventionist economists argue that the policies are necessary if post-tax income inequalities and poverty are to be reduced.

Questions

1. Describe how inequalities in the distribution of income and wealth have changed in recent years.
2. Evaluate the view that progressive taxation and transfers should not be used to reduce income inequalities and poverty.
3. Evaluate the view that the government should rely on market forces to determine the distribution of income and wealth in the UK, and should reduce direct intervention in the economy which has aimed at making the UK more equal.
4. Explain the difference between absolute and relative poverty.
5. Do you agree that faster economic growth which is needed to reduce poverty necessarily requires that the distribution of income becomes less equal? Justify your answer.



8

The market mechanism, market failure and government intervention in markets

Earlier chapters have explained how markets work. This chapter adds another dimension by introducing and explaining the four functions that prices perform in a market economy, or in the market sector of a mixed economy. It then explains how, when one or more of these four functions breaks down, market failure occurs.

The next sections look in some detail at the main forms of market failure: public goods, externalities, and merit and demerit goods. Externalities and merit and demerit goods are analysed with the help of the concepts of marginal private and external cost and benefit. The chapter also looks at environmental externalities, property rights and information failures associated with market failure. We then explain how income and wealth inequalities are a form of market failure.

In developed economies such as in the UK, the existence of market failure leads to government intervention in the economy to make markets function better, and in some cases to replace the market with state provision of goods and services. The last parts of the chapter introduce some 'industrial policy' issues. These are government competition policy, public ownership versus the privatisation of industries and services, and the regulation and deregulation of markets.

LEARNING OBJECTIVES

These are to understand:

- the four functions that prices perform and how markets and prices affect resource allocation
- the meaning of market failure
- the difference between a private good and a public good
- the difference between a merit good and a demerit good
- how externalities lead to market failure
- the concepts of marginal private, external and social cost and benefit
- how these concepts are used to analyse externalities, and merit and demerit goods
- how to apply the concept of allocative efficiency to market failure
- a number of environmental market failures
- the importance of property rights in the public policy response to externalities
- merit and demerit goods in the context of information failures
- how monopoly may lead to market failure and resource misallocation
- how income and wealth inequalities lead to market failure
- the ways government intervention in the economy tries to correct market failure and why it may fail.

8.1 How markets and prices allocate resources

The functions of prices

Earlier chapters have provided lots of information about the role of prices in a market economy or in the market sector of a mixed economy. So far, however, we have not drawn attention to the precise functions that prices perform. These are:

- the signalling function
- the incentive function
- the rationing function
- the allocative function.

Prices signal information

Prices provide information that allows buyers and sellers in a market to plan and coordinate their economic activities. This is the **signalling function of prices**. Here is an example. Most Friday afternoons, a man visits a local street market to buy fruit and vegetables for his family. He wants to buy both strawberries and raspberries. On reaching the market, he looks at the prices displayed by the market vendors on white tabs stuck into each tray of produce. The prices charged, which may not be the same at all the market stalls, provide vital information about what to buy. For example, the prices of strawberries and raspberries may signal the relative quality and scarcity of the fruit.

KEY TERMS

signalling function of prices prices provide information to buyers and sellers.

incentive function of prices prices create incentives for people to alter their economic behaviour, e.g. a higher price creates an incentive for firms to supply more of a good or service.

rationing function of prices rising prices ration demand for a product.

Prices create incentives

The information signalled by *relative* prices, such as the price of strawberries relative to the price of raspberries, creates incentives for people to alter their economic behaviour. A higher price in a market creates incentives for producers to supply more of a good or service because they believe that larger profits can be made. As Figure 8.1 shows, the incentive function of a price rise is shown by an extension of supply along the market supply curve, *S*, depicted by a movement up the left-hand arrow in the diagram. Likewise, falling prices reduce profits and thereby reduce the incentive to supply the product. Either way, this is the **incentive function of prices** in operation. Similarly, in labour markets, rising wages create the incentive for people to acquire new skills and supply their labour services, while falling wages reduce these incentives.

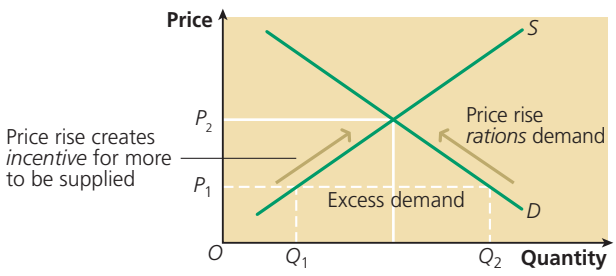


Figure 8.1 The incentive and rationing functions of prices

The rationing function of prices

Figure 8.1 also illustrates the **rationing function of prices**. At price P_1, the price rises to get rid of the excess demand shown in the diagram. The rising price, reflecting the strength of consumer preferences and consumers' ability to pay, rations demand for the good. The rationing function of prices is depicted by the movement up the right-hand arrow in the diagram, which shows a contraction of demand along the demand curve *D*. Demand contracts until excess demand is eliminated. In a similar way, rising wages in a labour market ration firms' demand for labour.

KEY TERM
allocative function of prices changing relative prices allocate scarce resources away from markets exhibiting excess supply and into markets in which there is excess demand.

Prices and resource allocation

The rationing function of prices is related to, but not quite the same as, the **allocative function of prices**. The rationing function distributes scarce goods to those consumers who value them most highly. By contrast, the allocative function directs resources between markets, away from the markets in which there is excess supply, towards the markets where there is excess demand. For example, a high price in the market for electric bicycles will encourage firms to allocate their productive resources (factors of production) in that market.

STUDY TIP
Make sure you understand and can explain, giving examples in particular markets, the signalling, incentive, rationing and allocative functions of prices. These functions are key to understanding the working of the price mechanism which you will be required to explain in some exam questions.

How prices coordinate the decision making of buyers and sellers

If you refer back to Application of Economics in the Real World 5.1, you will see a brief reference to the *invisible hand of the market*, an idea first used by the economist Adam Smith. Smith believed that if markets are highly competitive, approximating to what is now called perfect competition, both firms and consumers will passively accept the market prices set by the interaction of supply and demand in the market as a whole.

In a pure market economy, prices and markets alone perform the central economic task of allocating scarce resources between competing uses, but markets are also important, and usually dominant, in mixed economies. In both these types of economy, prices coordinate the decision making of buyers and sellers through their responses to the incentive function of prices which we have described, and then on through the rationing and allocative functions of prices.

Producers use the most efficient methods of production available to them to cut costs in pursuit of their business objective of profit maximisation. Consumers buy from sellers who charge the lowest price. Firms also switch productive resources into markets that maximise their return. All this takes place automatically through the invisible hand of the market.

STUDY TIP
Remember, of course, that in real-world markets, barriers to market entry and exit, and imperfect information, may distort the operation of Adam Smith's invisible hand of the market. Markets may therefore fail to achieve an efficient allocation of scarce resources.

The advantages and disadvantages of the price mechanism

Advantages

- Arguably, the operation of the price mechanism in competitive markets has the advantage of promoting consumer sovereignty, in the sense that the goods and services produced are those that consumers have voted for when spending the pounds in their pockets. When consumer sovereignty exists, the consumer holds the most power in the market. Firms and industries that produce goods other than those for which consumers are prepared to pay, do not survive in highly competitive markets.
- Providing it operates in this way, through cost reduction the price mechanism leads to a productively efficient allocation of resources.
- By redistributing resources into the production of goods and services that people wish to buy, the price mechanism achieves an allocatively efficient outcome.

Disadvantages

- As the name suggests, imperfectly competitive markets are characterised by asymmetric market information and market power, both of which favour firms or producers rather than consumers. In extreme cases, as in monopoly, the operation of the price mechanism leads to an outcome in which firms exploit their producer sovereignty. The goods and services available for consumers to buy are determined by the firms rather than by consumer preferences expressed in the marketplace.
- Even if producer sovereignty is not exercised on a 'take-it-or-leave-it' basis by a monopoly, the highly imperfectly competitive firms may still possess sufficient market power to manipulate consumer wants through such marketing devices as persuasive advertising. In these situations, the producer rather than the consumer holds the most power in the market.
- The price mechanism is 'value neutral' in that it has no regard for equality and the fairness or otherwise of the allocation of buying power between different income groups. And as we will see in the next section, the unrestricted operation of the price mechanism may lead to a number of significant market failures.

The case for and against extending the operation of the price mechanism into new areas of activity

Pro-free-market economists believe that markets and the price mechanism work well and that government intervention in the economy works badly. Many also take the view that, because of the risk of government failure which produces an outcome worse than market failure, the price mechanism should be extended into parts of the economy previously dominated by state provision of goods and services and the operation of the command mechanism or planning mechanism.

However, other economists, whom we will call interventionist economists, take the opposite view, believing that markets often perform badly and that government intervention and the planning mechanism can often improve on the free operation of the price mechanism. We will go deeper into this debate in section 8.10 on government failure.

SECTION 8.1 SUMMARY

- Prices perform four functions in markets: signalling information, creating economic incentives, rationing scarce resources and allocating them between competing uses.
- Prices signal information that allows buyers and sellers in a market to plan and coordinate their economic activities.
- The information signalled by *relative* prices creates incentives for people to alter their economic behaviour.
- Rising prices ration demand for a product.
- Changing relative prices allocate scarce resources away from markets exhibiting excess supply, and into markets in which there is excess demand.
- Prices coordinate the decision making of buyers and sellers through their responses to the incentive function of prices, and then on through the rationing and allocative functions of prices.
- The operation of the price mechanism in competitive markets has the advantage of promoting consumer sovereignty.
- When consumer sovereignty exists, the consumer holds the most power in the market. Firms and industries that produce goods other than those for which consumers are prepared to pay, do not survive in highly competitive markets.
- In imperfectly competitive markets and monopoly, the operation of the price mechanism leads to an outcome in which firms exploit their producer sovereignty. In these situations, the producer holds the most power in a market.
- The price mechanism is 'value neutral' in that it has no regard for equality and the fairness or otherwise of the allocation of buying power between different income groups.

KEY TERMS

market failure when the market mechanism leads to a misallocation of resources in the economy, either completely failing to provide a good or service or providing the wrong quantity.

complete market failure a market fails to function at all and a 'missing market' results.

partial market failure a market does function, but it delivers the 'wrong' quantity of a good or service, which results in resource misallocation.

missing market a situation in which there is no market because the functions of prices have broken down.

8.2 The meaning of market failure

Market failure occurs whenever a market leads to a misallocation of resources. It occurs whenever the market mechanism or price mechanism performs badly or unsatisfactorily, or fails to perform at all. It is useful to distinguish between **complete market failure**, when the market simply does not exist, and **partial market failure**, when the market functions but produces the 'wrong' quantity of a good or service. In the former case, there is a '**missing market**'. In the latter case, the good or service may be provided too cheaply, in which case it is over-produced and over-consumed. Alternatively, as in monopoly, the good may be too expensive, in which case under-production and under-consumption result.

Relating market failure to the functions of prices

In section 8.1, when explaining how the price mechanism distributes scarce resources between alternative uses in a market or mixed economy, we introduced the four functions that prices perform in such economies. To recap, these are: signalling information; creating incentives to influence people's

behaviour; allocating scarce resources between competing uses; and rationing the demand for goods and services.

Ignoring the possibility that market failure can be associated with inequalities in the distribution of income, we can say that, when all four of these functions perform well, markets also work well and market failure is either non-existent or trivial. However, when one or more of the four functions of prices significantly breaks down, market failure occurs. You will see in sections 8.3 and 8.4 how, in the case of pure public goods and externalities, the price mechanism breaks down completely. If an alternative method of provision does not exist, none of the public good is produced and there is complete market failure. In the case of an externality such as pollution, firms (and indeed consumers) that generate pollution simply dump it on other people (whom we call third parties). There is no market in which the unwilling consumers of pollution can charge producers for the discomfort they suffer. The lack of a market means there is no incentive for the polluter to pollute less. Hence, again, there is market failure.

TEST YOURSELF 8.1
Do you agree that a highly unequal distribution of income is a market failure? Justify your answer.

SECTION 8.2 SUMMARY

- Market failure occurs when the market mechanism leads to a misallocation of resources in the economy, either completely failing to provide a good or service or providing the wrong quantity.
- Complete market failure occurs when a market fails to function at all and a 'missing market' results.
- When all four of the functions of prices perform well, markets also work well and market failure is either non-existent or trivial.
- However, when one or more of the four functions of prices significantly breaks down, market failure occurs.

8.3 Private goods, public goods and quasi-public goods

Private goods

Most goods are **private goods**, which possess two defining characteristics.

- The owners can exercise private property rights, preventing other people from using the good or consuming its benefits. A pure private good is **excludable** — for example, a shopkeeper can prevent people from consuming the goods on display in her shop, unless they are prepared to pay for them. (Shoplifters, of course, try to get round this exercise of private property rights!)
- The second property of a pure private good is that it is **rival**. Rivalry can be illustrated in the context of a good such as chocolate. If you eat a bar of chocolate, other people cannot eat it and gain its benefits. In this sense, people are rivals. (Rivalry is sometimes called diminishability. When one person consumes a private good, such as a sweet or a banana, the quantity available to others diminishes.)

KEY TERMS
private good a good, such as an orange, that is excludable and rival.
excludable good people who do not want to pay can be excluded from benefiting from the good.
rival good when one person consumes a private good, the quantity available to others diminishes.

Public goods

KEY TERM
public good a good, such as a radio programme, that is non-excludable and non-rival.

A **public good** exhibits opposite characteristics to those of a private good, as it is non-excludable and non-rival. It is these characteristics that lead to market failure.

A lighthouse, or rather the beam of light provided by a lighthouse, is an example of a public good. Suppose an entrepreneur builds the lighthouse shown in Figure 8.2, and then tries to charge each ship that passes in the night and benefits from the beam of light. As long as ships pay up, the service can be provided commercially through the market.

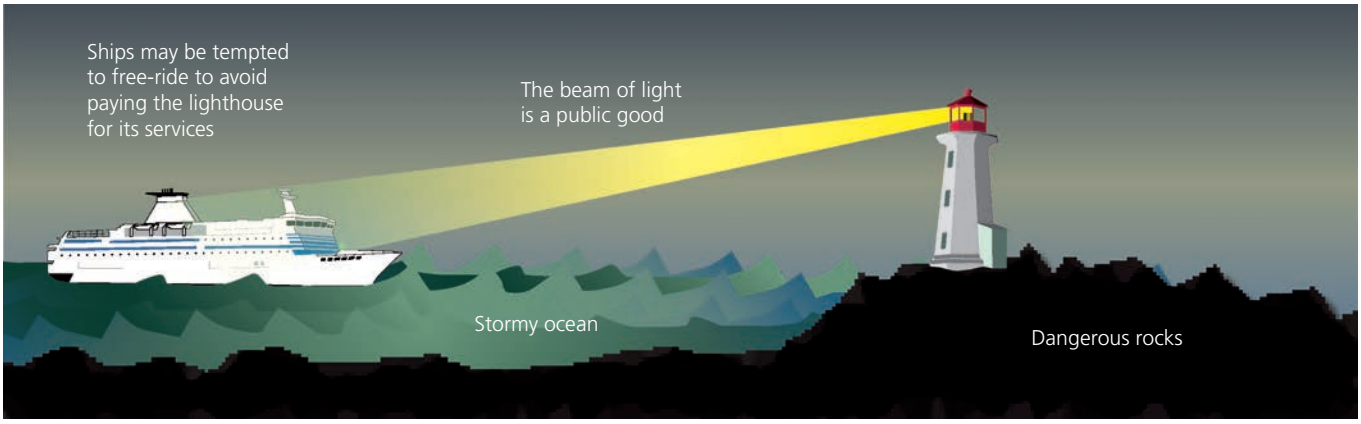


Figure 8.2 The beam of light provided by a lighthouse is a public good

However, the market is likely to fail. The appearance of one more ship near the lighthouse does not prevent other ships from sailing in the area. Non-rivalry thus occurs. It is also impossible to exclude free-riders (in this case, ships that benefit without paying). Because of non-excludability, it may be impossible to collect enough revenue to cover costs. If too many ships decide to 'free-ride', profits cannot be made and the incentive to provide the service through the market disappears. The market thus fails to provide a service for which there is an obvious need. There is then a need for alternative provision by the government in its public spending programme, or possibly by a charity (such as Trinity House in the UK — a charity safeguarding shipping and seafarers).

STUDY TIP
Make sure you understand the difference between a private good and a public good, and can give examples of both, if required, in your exam. Their characteristics mean that they are the opposite of one another.

TEST YOURSELF 8.2
Should governments in all circumstances provide public goods? Justify your answer.

Other examples of public goods

Other traditional examples of public goods include national defence, police, street lighting, roads, and television and radio programmes. To take one of these, national defence is non-rival in the sense that receiving the benefits of national defence (e.g. greater peace of mind) does not reduce the benefits available to other people. National defence is also non-excludable because providing the benefits for one means providing the benefits for all.

To explain this further, consider a situation in which the state does not provide national defence. Instead, the government lets individual citizens purchase in the market the defence or protection they want. But markets only provide defence when entrepreneurs can successfully charge prices for the services they supply. Suppose an aspiring citizen, who believes a fortune can be made in the defence industry, sets up a company, Nuclear Defence Services Ltd, with the aim of persuading the country's residents to purchase the services of nuclear missiles strategically located around the country. After estimating the money value of the defence received by each individual, Nuclear Defence Services Ltd bills each household accordingly and waits for the payments to flow in.

This is clearly not something the general public would agree to, and the payments may never arrive. As long as the service is provided, every household can benefit without paying. Nuclear Defence Services Ltd cannot provide nuclear defence only to the country's inhabitants who are prepared to pay, while excluding the benefit from those who are not prepared to pay. Withdrawing the benefit from one means withdrawing it from all. But all individuals face the temptation to consume without paying, or to free-ride. If enough people choose to free-ride, Nuclear Defence Services Ltd makes a loss. The incentive to provide the service through the market thus disappears. Assuming, of course, that the majority of the country's inhabitants believe nuclear defence to be necessary (i.e. a good rather than a 'bad'), the market fails because it fails to provide a service for which there is a need. The result is a missing market.

TEST YOURSELF 8.3

Explain why D is the correct answer to the following question, and why A, B and C are incorrect.

When applied to a public good, 'non-rival' means that:

A there is a single monopoly supplier of the good

B the resources used in the good's production could not have been used to produce other goods

C if the good is provided for one person, it must be provided for everybody else

D consumption of the good by one person does not reduce the amount of the good available to others

EXTENSION MATERIAL

'Goods' and 'bads'

A good such as a loaf of bread provides benefits to the person or persons who consume it. Consumer goods yield usefulness or utility, and sometimes pleasure and satisfaction. (Using economic jargon, we say that the consumption of consumer goods increases economic welfare.)

In everyday language, we generally use the word 'bad' as an adjective: for example, a bad film or a bad football match. However, economists also use the word as a noun. In this usage, an economic 'bad' is the opposite of a good, yielding disutility, dissatisfaction or displeasure. For most people, consumption of a bad such as rotten meat reduces rather than increases economic welfare.

Likewise, a 'public bad' is the opposite of a public good. An example of a public bad is rubbish or garbage. The production of public bads such as garbage leads to a free-rider problem, though the problem is subtly different from the one that occurs in the case of public goods. People are generally prepared to pay for the removal of garbage, to avoid the unpleasantness otherwise experienced. However, payment can be avoided by dumping the bad in a public place or on someone else's property.

In the UK, local authorities generally empty household dustbins without charging for each bin emptied. Suppose this service is not provided, and private contractors remove rubbish and charge households £1 for each dustbin emptied. To avoid paying £1, some households may decide to dump their waste in the road or in neighbours' dustbins. (Builders' skips provide a good example of this practice. A household hiring a skip is well advised to fill the skip as quickly as possible, before the rest of the street takes advantage of the facility.) If too many households free-ride, it is impossible for the private contractor to make a profit, and a service for which there is a need is no longer provided. Hence the case for free local authority provision, financed through taxation.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.1

The Seattle stomp

For several years some US local authorities have been charging households for emptying their dustbins, arguing that if rubbish disposal is free, people produce too much rubbish. The obvious economic solution is to make households pay the cost of disposing of their waste. That will give them an incentive to throw out less and recycle more (assuming that local governments provide collection points for suitable materials). Several US cities have started charging households for generating rubbish. The common system is to sell tags which householders attach to rubbish bags or bins. In effect, the price of a tag is the price the household pays for creating another bag of rubbish.

In Charlottesville, Virginia, households were charged 80 cents for each tag they bought. Following the introduction of pricing, the number of rubbish bags collected fell sharply, by 37% over the first five months. However, this was largely due to a practice nicknamed the 'Seattle stomp' — a frantic dance first noticed when Seattle introduced rubbish pricing. Rather than buy more tags, households simply crammed 40% more rubbish into each bag. But this is inefficient as compacting or crushing is better done by machines at landfill sites than by households. The weight of rubbish collected at Charlottesville (a better indicator of disposal costs than volume) fell by a modest 14%, while in nearby cities with no pricing scheme, it fell by 3.5%.

More significantly, people resorted to illegal dumping or fly-tipping rather than paying to have their rubbish removed. Fly-tipping may have accounted for 30–40% of the reduction in collected rubbish. There was a 15% increase in the weight of materials recycled, suggesting that people chose to recycle free rather than pay to have their refuse carted away. But cities with rubbish-pricing policies are likely to have 'greener' citizens who recycle more in any case. Once this effect is removed, it appears that pricing rubbish collection has no significant effect on recycling. Also, people tend to throw out more rubbish in richer towns than in poorer ones.

Follow-up questions

1. Why do most local councils in the UK empty household dustbins without charging a price for the service?
2. Why do most local councils in the UK charge a price for emptying household garden waste bins?
3. Evaluate the case for charging a price for emptying dustbins.
4. What is meant by fly-tipping?
5. What policies might a behavioural economist recommend to reduce fly-tipping?

Quasi-public goods

KEY TERM
quasi-public good a good which is not fully non-rival and/or where it is possible to exclude people from consuming the product.

Public goods can be divided into pure public goods and **quasi-public goods**. National defence and police are examples of pure public goods — defined as public goods for which it is impossible to exclude free-riders. However, most public goods (street lighting, roads, television and radio programmes, and also lighthouses) are really quasi-public goods (also known as non-pure public goods).

Methods can be devised for converting the goods into private goods by excluding free-riders (e.g. electronic pricing of road use). Quasi-public goods can be provided by markets, but the second property of non-rivalry or non-diminishability means there is a case for providing all public goods free in order to encourage as much consumption as possible. If we assume the lack of a capacity restraint, for public goods and quasi-public goods, the optimal level of consumption occurs when they are available free of charge. (However, capacity constraints, resulting from factors such as the limited ability of roads to carry cars and the limited bandwidth of airwaves, mean that there is also a case for charging for public goods such as roads and airwaves to deal with the problem of congestion.)

TEST YOURSELF 8.4
Using an example, explain the difference between a pure public good and a quasi-public good.

The significance of technological change

Roads provide one of the best examples of a good which has changed over the years from being regarded as a pure public good to now being thought of as a quasi-public good. Until quite recently, except in the case of motorways where limited points of access meant that toll gates could be installed, it was deemed impractical to charge motorists for the use of ordinary roads.

This has now changed. Technological change makes it feasible for the government or local authorities to use electronic pricing to charge *all* motorists for road use. The charges can also be varied at different times of day to create incentives for car drivers to shift their travel patterns from rush hour to non-rush hour. There is a case for *not* charging a price when roads are uncongested and thus largely non-rival, but for charging for road use as soon as the problem of congestion and road rivalry occurs. So why are there so few road pricing schemes? After all, we are happy to pay for most goods and to pay different prices at different times of day: for example, for watching a movie at a cinema. The answer to this question probably lies in the political power of the motoring lobby. Politicians fear they will not get re-elected if they try to bring in road pricing schemes.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.2

Twenty years of the London congestion charge

In February 2023, the London congestion charge was 20 years old. At the time of its introduction, many economists believed that the congestion charge marked the triumph of economic common sense over narrow self-interest. Economists confidently believed that many other cities, both in the UK and abroad, would rush to adopt London-style road pricing. In the event, this did not happen. In November 2008, Manchester voters were asked to approve the introduction of a city-centre congestion charge and they voted against it.

What have been the possible benefits and costs of the London congestion charge? In 2003 the idea of charging car users to drive around the capital was met with near-apocalyptic warnings from motoring groups and newspapers such as the *Daily Mail*. The charge would 'destroy' the city's commercial heart and cause 'total gridlock', warned some. It would



The London congestion charge has reduced car use but increased the use of taxis and private hire vehicles

'cause misery to thousands of commuters across the capital', said Conservative Greater London Authority's transport spokesman, Angie Bray. 'Londoners will suffer conditions worse than cattle trucks on their morning commute into work,' she added.

Since it was introduced in 2003, however, the charge, which has risen from £5 to £15, has not caused gridlock. Rather, it has resulted in a slight reduction in traffic levels. But while car numbers are down, the number of private hire vehicles — minicabs and Ubers — is up. Trips by taxi and private hire vehicles as the main mode of the journey increased by 9.8% between 2015 and 2016 alone — and by 29.2% since 2000. Today, more than 18,000 different private hire vehicles enter the congestion charging zone each day, with peaks on Friday and Saturday nights. This has reduced the speed of traffic through the city centre, which in turn has affected the bus network.

Follow-up questions

1 Why have other cities, in the UK and the rest of the world, not rushed to follow London's example and introduce a congestion charge?
2 What policies have cities in other countries introduced to try to reduce road congestion?
3 Do you agree that road use should be priced? Justify your answer.
4 It has been said that the real aim of the London congestion charge is to reduce air pollution rather than traffic congestion: for example, by giving exemption to electrically powered cars. Do you think these two objectives are compatible? Justify your reasoning.

SECTION 8.3 SUMMARY

- A private good is a good or service, such as an apple, that is excludable and rival.
- Excludability occurs when people who are unprepared to pay can be excluded from benefiting from a good or service.
- Rivalry occurs when one person consumes a private good or service and the quantity available to others diminishes.
- A public good is a good or service, such as a radio programme, that is non-excludable and non-rival.
- Examples of public goods include national defence, police, street lighting, roads, and terrestrial television and radio programmes.
- Public goods can be divided into pure public goods and quasi-public goods.
- Roads provide an example of a good which has changed over the years from being regarded as a pure public good to now being thought of as a quasi-public good.
- For public goods and quasi-public goods, the optimal level of consumption occurs when they are available free of charge — provided there is no capacity constraint.

8.4 Positive and negative externalities in production and consumption

What are externalities?

Externalities exist when there is a divergence between private and social costs and benefits. Essentially, an **externality** is a special type of public good or public 'bad' which is 'dumped' by those who produce it onto other people (third parties) who receive or consume it, whether or not they choose to. The key feature of an externality is that there is no market in which it can be bought or sold — externalities are produced and received outside the market, providing another example of a missing market.

KEY TERMS

externality a public good, in the case of an external benefit, or a public bad, in the case of an external cost, that is 'dumped' on third parties outside the market.

positive externality an external benefit that occurs when the consumption or production of a good causes a benefit to a third party, where the social benefit is greater than the private benefit.

negative externality an external cost that occurs when the consumption or production of a good causes costs to a third party, where the social cost is greater than the private cost.

property right the exclusive authority to determine how a resource is used.

Positive and negative externalities

The provider of an external *benefit* (or **positive externality**), such as a beautiful view, cannot charge a market price to any willing free-riders who enjoy it, while conversely, the unwilling free-riders who receive or consume external *costs* (or **negative externalities**), such as pollution and noise, cannot charge a price to the polluter for the bad they reluctantly consume.

Externalities and property rights

A **property right** is the exclusive authority to determine how a resource is used, whether that resource is owned by government or by individuals. The owner of a house, for example, has the right to live in the house, to let others such as the members of their family live in the house without paying rent, to rent the property to a tenant, or to sell the house to a new owner. In modern capitalist economies, the legal system gives house owners the power to exercise private property rights in one or other of these ways.

However, when externalities are generated, the right to exercise private property rights can disappear. For example, the owner of a beautiful house which is visible from the public highway cannot prevent passers-by from enjoying the positive externality of the view of the house. The owner might be able to build a high wall to obstruct this view, or even to destroy the house, but in most situations, passers-by can free-ride: in other words, enjoy the benefits without paying.

KEY TERM

free-rider problem a free-rider is someone who benefits without paying as a result of non-excludability. Customers may choose not to pay for a good, preferring instead to free-ride, with the result that the incentive to provide the good through the market disappears.

The free-rider problem

Earlier, we briefly mentioned the **free-rider problem** in the extension material on 'goods' and 'bads'. We have seen that the owner of a pure public good cannot exclude people who are unprepared to pay from gaining the benefits of the good. These people can free-ride, gaining the benefits without paying. Likewise, the provider of an external *benefit* (or positive externality), such as a beautiful view, cannot charge a market price to any willing free-riders who enjoy it, while conversely, the unwilling free-riders who receive or consume external *costs* (or negative externalities), such as pollution and noise, cannot charge a price to the polluter for the bad they reluctantly consume. With both public goods and externalities, the free-rider problem is a cause of market failure.

The free-rider problem occurs when non-excludability leads to a situation in which not enough customers choose to pay for a good, preferring instead to free-ride, with the result that the incentive to provide the good through the market disappears and a missing market may result.

Externalities can be generated by firms in the course of producing the goods they eventually sell in a market, or they can be generated by households and individuals when they consume goods. The former are known as *production* externalities, while the latter are *consumption* externalities. We will look at these in turn.

STUDY TIP

Make sure you understand the difference between negative and positive externalities, and between production and consumption externalities, and can give examples of all of these. Many of the examples of market failure you will come across in your studies fall into one, or several, of these categories.

Production externalities

Negative production externalities

KEY TERM

production externality an externality (which may be positive or negative) generated in the course of producing a good or service.

Consider the power station illustrated in Figure 8.3, which discharges pollution into the atmosphere in the course of producing electricity. We can view a negative **production externality** (or external cost) such as pollution as being that part of the true or real costs of production which the power station evades by dumping the bad on others: for example, the people living in the houses and the businesses in the commercial forestry industry. The price that the consumer pays for the good (electricity) reflects only the money costs of production, and not all the real costs, which include the external costs (including the eyesore or visual pollution also shown in the diagram). In a market situation, the power station's output of electricity is thus under-priced. The allocative function of prices has once again broken down — under-pricing encourages too much consumption of electricity, and therefore over-production of both electricity and the spin-off, pollution.

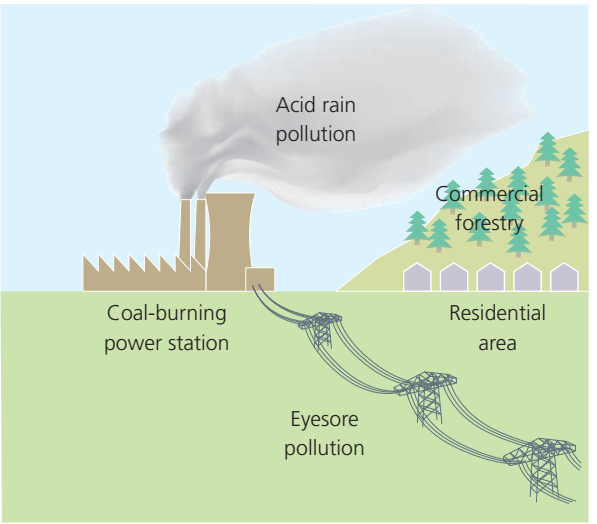


Figure 8.3 The discharge of negative production externalities by a power station

Positive production externalities

Figure 8.4 shows again the power station previously illustrated in Figure 8.3, but in this case the production of electricity yields positive production externalities (or external benefits) rather than negative externalities. We have assumed that the power station discharges warm (but clean) water into the lake adjacent to the power station. Warmer temperatures increase fish stocks, and commercial fishing boats and private anglers then benefit. Unless it owns the lake, the power station company cannot charge the fishers for the benefits they are receiving. (You might, of course, query our assumption that the water discharge creates positive rather than negative externalities. In all likelihood, disruption of a local ecosystem might cause negative externalities, such as algae pollution.)

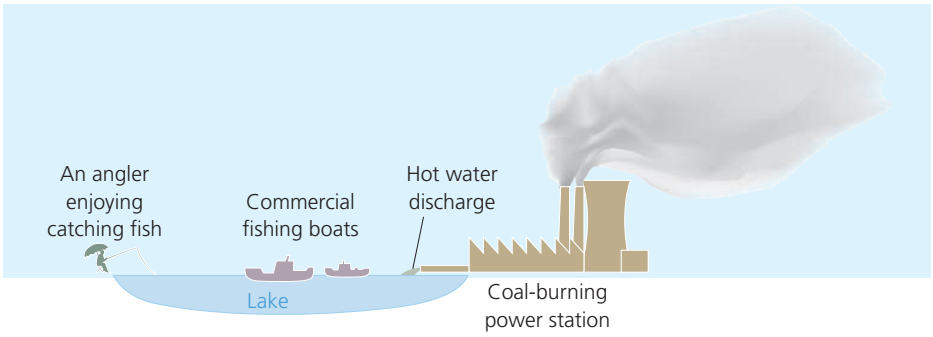


Figure 8.4 The discharge of positive production externalities by a power station

Consumption externalities

Negative consumption externalities

KEY TERM
consumption externality an externality (which may be positive or negative) generated in the course of consuming a good or service.

We are all probably familiar with the annoying experience of going to the cinema and having our pleasure disrupted by the ringing of mobile phones or the noisy eating of popcorn by the person sitting a few seats away. These are examples of negative **consumption externalities**, unwillingly received by a cinema goer as a result of consumption activities undertaken by other members of the audience. There are many examples of similarly annoying externalities — walking through litter and chewing gum dropped by pedestrians, being splashed by inconsiderate motorists, and having to listen to a neighbour's loud music regularly played through an open window.

Positive consumption externalities

A good example of a positive consumption externality is the pleasure gained by a passer-by walking past beautiful buildings and household gardens in a residential area. At this point, it is also worth noting that a negative consumption externality suffered by one person may be a positive consumption externality for somebody else.

Consider also a situation when a person pays to watch a Premier League football match. The football fan would gain less pleasure from watching the match if the stadium were empty than if soaking up the 'atmosphere' generated by other fans in a packed stadium. Much the same happens in restaurants. Prospective diners walk past empty or nearly empty restaurants, preferring instead to queue for a table at an already full eatery. Maybe they think the food is better, but despite having to wait longer for their food, what they also enjoy is the 'atmosphere' generated by other diners.

Summary: different types of externality

As we have explained, externalities divide into negative externalities and positive externalities (which are also known as external costs and benefits). Table 8.1 lists the possible types of externality and gives examples of negative externalities (costs) and positive externalities (benefits) for each.

Table 8.1 Examples of the different types of externality

Type of externality	Negative externalities	Positive externalities
Pure production externalities *(generated and received in production)*	Acid rain pollution discharged by a power station which harms a nearby commercially run forest	A farmer benefiting from drainage undertaken by a neighbouring farmer
Pure consumption externalities *(generated and received in consumption)*	Noisy music at a party disturbing neighbouring households	Households benefiting from the beauty of neighbouring gardens

How externalities lead to the 'wrong' quantity of a good being produced and consumed

We mentioned earlier that when negative externalities are generated in the course of production, part of the true or real costs of production are not borne by the producer, instead being dumped on others. This means that, even in competitive markets, if *negative production* externalities are generated, goods end up being too cheap or under-priced. The market has created the wrong incentives. As prices under-reflect the true costs of production, which include the cost of the negative externalities, too much of the good ends up being produced and consumed.

The opposite happens when firms generate *positive production* externalities. Prices end up being too high, leading to the 'wrong' quantity of the good being produced and consumed. Once again, the market has created the wrong incentives, which discourages consumption. In this case, prices over-reflect the true costs of production. Not enough of the good ends up being produced and consumed.

In much the same way, when *negative* and *positive consumption* externalities are generated, markets are distorted and the wrong quantity of a good is produced and consumed. This is explained in section 8.5 on merit and demerit goods.

STUDY TIP

Students often fail to understand that externalities are generated and received outside the market. Remember that both public goods and externalities provide examples of missing markets.

EXTENSION MATERIAL

Establishing markets for trading private property rights

In 1960, Professor Ronald Coase argued that, if markets can be created for private property rights, government intervention to correct market failures may not be necessary. Coase used the example of wood-burning locomotives, which in nineteenth-century America frequently set fire to farmers' fields. If farmers possess the property right to prevent crops being destroyed, they can sell the right to the railway companies. By contrast, if the railway companies possess the property right to emit sparks, farmers could pay the companies to reduce emissions.

Coase argued that in both cases the outcome might be the same. If farmers have a right to stop the sparks, but emitting sparks is worth more to the railway than stopping the sparks is worth to the farmers, the railway will buy the right to emit sparks from the farmers, and the damage continues. But if the railway companies have the right to emit sparks, and this right is worth more to them than to the farmers, the right will not be sold, and the damage again continues. In this example, initial ownership of property rights has no effect on the amount of resources devoted to suppressing sparks. The ability to trade property rights ensures the same outcome in either case.

This theory, which is known as the Coase theorem, has greatly influenced the free-market approach to market failures. Indeed, most economists now accept that governments should try to work *with the market* rather than *against the market* through regulation.

The tragedy of the commons



Garrett Hardin, who developed the theory of 'the tragedy of the commons'

Critics of the private ownership of property rights do not generally want to abolish those rights. Instead, they wish to transfer property rights from private ownership to government ownership. However, the worst-case scenario is when private property rights are abolished and replaced by 'common ownership', but without any form of government control that aims to achieve the 'common good'.

In 1968, in an article titled 'The tragedy of the commons', Garrett Hardin told the following story (we have adapted Hardin's story into our own words):

> In Britain before the eighteenth century, much farmland was open common land on which farmers could graze as many animals as they wished, for free. As long as poaching and disease keep the numbers of animals below the carrying capacity of the land, common land grazing works well. However, it was in a herdsman's self-interest to graze more and more animals, since his gain, the money earned from selling slaughtered additional animals, would exceed any loss incurred from the gradual deterioration of the common. The herdsman would receive all the proceeds from the sale of an additional animal, but the effects of over-grazing would be shared by all the herdsmen.
>
> A rational herdsman would therefore conclude that the only sensible course for him to pursue was to add another animal to his herd. But each herdsman would then be locked into a system that compelled him to increase his herd without limit — in a world that is limited. Another animal would be added to the grazing land, and then another, and then another … But this would be the conclusion reached by each and every rational herdsman sharing the common. Therein is the tragedy. Freedom of the commons brings ruin to all.

A few years later in 1974, Hardin used the following example to illustrate the tragedy of the commons. Satellite photos of northern Africa showed an irregular dark patch over an area of 390 square miles. Ground-level

investigation revealed a fenced area inside which there was plenty of grass. Outside, the ground cover had been devastated. Hardin argued that the explanation was simple. The fenced area was private property, subdivided into sections. Each year, the owners moved their animals to a new section. Fallow periods of four years gave the pastures time to recover from the grazing. The owners did this because they had an incentive to take care of their land. But no one owned the now barren land outside the fenced-off area. This land being open to nomads and their herds, the tragedy of the commons was taking its toll.

Elinor Ostrom was the first female economist to win the Nobel Memorial Prize in Economics (in 2009), for her work on common resources. Ostrom's research challenged Hardin's ideas, arguing that individuals and groups could in fact manage their collective resources very effectively — a non-tragedy of the commons. Her work, carried out in Indonesia, Kenya and Nepal, concluded that resources such as land used for growing crops and grazing animals are managed best when those who benefit from their use live in close proximity to the resources. She further argued that regulation of common resource usage should be administered at local community level, rather than national level, since people are more likely to monitor, adhere to and enforce rules they have had an input in setting. There is some overlap here with the behavioural economics principles of altruism and social norms covered in Chapter 2.

SYNOPTIC LINK

Refer back to the section on 'Externalities and property rights' on page 235. The absence of clearly defined property rights can be a reason for market failure, including negative externalities.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.3

Has the fishing industry fallen victim to the tragedy of the commons?

For decades environmentalists have argued pressures on the Earth's natural resources are due to the increasing size of the human population. For some resources such as oil, the market mechanism has encouraged consumers to economise and producers to search for and to develop new sources of supply.

But with fish it may be different. Unless stocks are managed tightly by all concerned with fishing, they may well collapse, and soon another 'tragedy of the commons' will have been played out. It happens because people think they can take a limitless amount of the Earth's 'free gifts', such as the atmosphere or the sea, or now, we are realising, fish.

For centuries these apparently free goods had no prices attached to them and so there was nothing restraining their use. And for centuries, nothing harmful happened. But on a finite globe, the limits logically have to be reached at some stage. A fish stock can make what is known as an 'equilibrium shift'; it can change under pressure to a new level of stable numbers much lower than they were before. Intensive fishing pressure may take out many of the bigger fish so that breeding slows down, until virtually no new breeding stock is coming through.

In the early 2000s it was widely argued that the decline in the North Sea might be irreversible if very stringent measures were not taken. A further problem then detected was the advent of global warming: as the sea waters warm, cod seem to be moving further north to cooler waters where they prefer to spawn. But at the time, there was little doubt that fishing pressure was the main threat.

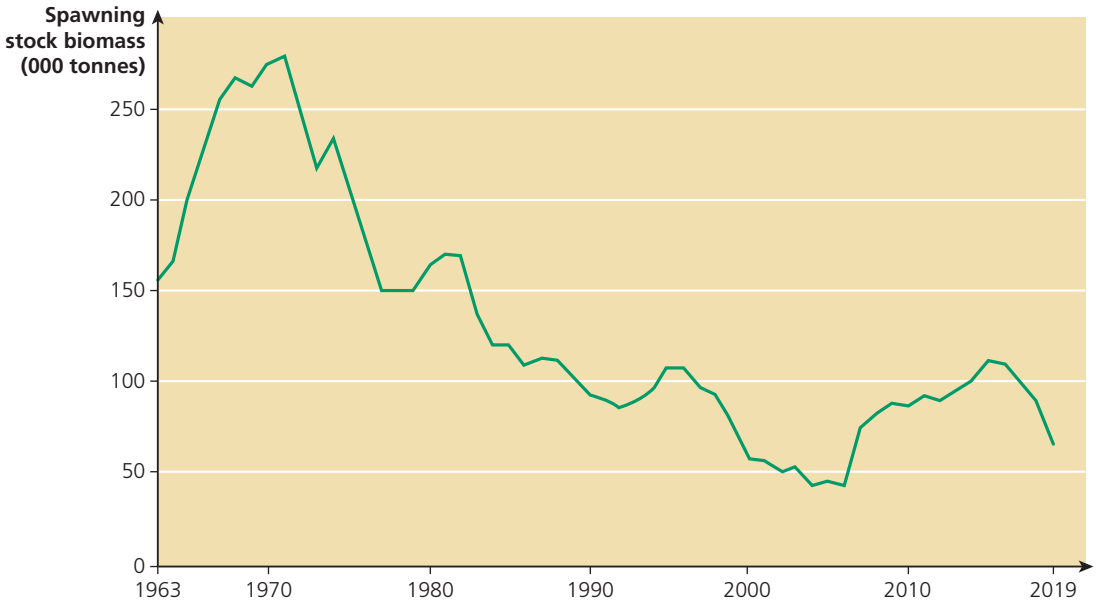


Figure 8.5 North Sea cod stocks, 1963–2019

Source: Marine Management Organisation

The graph shows that, although the spawning stock biomass (SSB) of North Sea cod (the fish stock population capable of reproducing) fell dramatically from around 1970 until 2006, the SSB then recovered and by 2015 had reached its highest level for 35 years. However, the SSB did fall significantly after 2016.

Follow-up questions

1 Do you agree that the depletion of North Sea cod stocks prior to 2006 was an example of the tragedy of the commons?
2 Find out whether stocks of North Sea cod have recovered up to the time you are reading this case study.
3 Research whether the recovery of North Sea cod stocks should be regarded as a success for the European Union's Common Fisheries Policy.
4 State two other environmental resources which might be suffering a tragedy of the commons.

Environmental market failures

Some of the market failures we have discussed result from the discharge of negative externalities into the environment. Such environmental externalities often involve pollution of land, sea, rivers and air, and the externalities associated with road use and congestion. We have described how coal-burning power stations discharge sulphur dioxide pollution into the atmosphere, and also the 'eyesore' pollution caused by unsightly buildings and electricity transmission lines. Later in the chapter we will explain how governments can attempt to correct such market failures through taxation, regulation including outright bans, or a combination of these approaches. Governments are also increasingly making use of behavioural 'nudges' to reduce environmental market failures such as those caused by the dumping of plastic into rivers and oceans. The A-level specification also emphasises that exam candidates should appreciate the relevance of the 'tragedy of the commons' for environmental market failures.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.4

The Green Deal

Between 2010 and 2013 the Behavioural Insights Team (BIT) worked with the Department of Energy and Climate Change (DECC) and the Department for Communities and Local Government (DCLG) to launch the government's flagship environmental policy: the Green Deal.

For years governments have offered subsidies on green products such as energy-saving light bulbs and loft installation. Traditional economic policies, which create incentives by reducing the price of products, have had limited success. The Behavioural Insights Team has used its understanding of the behavioural factors that determine household decision making to develop policies that encourage energy-saving behaviour among households and firms.

Choosing small rewards today rather than large rewards in the future

It is reasonable to assume that households will invest in energy-saving measures in their homes that reduce heating costs. New boilers, floor, wall and loft installation, double or triple glazing, energy-efficient external doors and other improvements insulate buildings so that they retain heat and bring down energy bills. However, many households decide against undertaking such spending because they are unwilling to pay large upfront costs today for benefits which will arrive in small instalments spread over many years. When left to themselves without any policy 'nudges' that might change their way of thinking, households prefer small rewards *today* rather than much larger rewards *in the future*.

The Green Deal was designed to change household behaviour. It did this by helping households pay for the upfront costs of investing in energy-efficiency measures. Private-sector companies, charities and local authorities paid for the upfront cost of energy-efficiency products and installation. Households then paid back these costs through their electricity bills. However, the repayment costs were not as high as they otherwise might have been because the households were benefiting from the energy-saving investments that reduced their electricity bills. The Green Deal rested on the assumption that, over the years ahead, household energy bills, which include the repayment costs, would not be higher than would have been the case if no energy-saving measures had been undertaken. In the long run, when the loan had been paid off, households would experience significantly lower energy bills and gain all the benefit.

The behavioural economic insight is that, by deferring the cost of energy-saving measures, the government could encourage individuals to invest in products that improve the efficiency of the nation's housing stock and reduce carbon emissions.

Evaluating the Green Deal

Demand for the Green Deal was exceptionally high. When the Green Deal Home Improvement Fund (GDHIF) was launched in June 2014, it spent its entire year's £120 million budget within six weeks. The GDHIF provided 20,000 households with vouchers to pay for energy-saving improvements on their homes. In December 2014 the DECC released a further £24 million which was taken up within 24 hours.

Consumer groups generally supported the Green Deal but they raised three main criticisms. First, a *Which?* magazine investigation found that the quality of the assessments conducted by Green Deal assessors varied widely. Some assessors produced good assessments of the energy-saving measures required in a property, but others failed to identify problems correctly and produced flawed reports. This was a systems failure that is likely to occur in any policy programme introduced by either public- or private-sector organisations.

Second, households who signed up for the Green Deal might eventually have to pay higher energy bills once the repayment costs had been added in.

Third, the Green Deal was extremely complex and charged a rate of interest of up to 10.8% APR (APR stands for annual percentage rate and is the annual cost of taking out a loan, including fees, expressed as a percentage of the sum borrowed). Households might have done better by paying for home improvements upfront themselves, or through personal loans.

The end of the Green Deal

In July 2015 the newly elected Conservative government announced the end of the Green Deal, which had been introduced with great fanfare by the previous Coalition government. The DECC said it had decided to end the Green Deal in order to protect taxpayers, citing low take-up and concerns about industry standards. The opposition Labour Party argued that government policy to improve energy efficiency had been a 'failure'. Roger Harrabin, the BBC environment analyst, said:

> The decision has caused anger — not because anyone really likes the Green Deal, but because it is being scrapped without a replacement in sight. The home insulation industry says it will create huge uncertainty for firms while the government works out what the replacement policy should be. There is no doubting the need for a scheme to finance improvements in home energy. The UK has some of the worst-insulated housing stock in Europe, and saving energy is by far the cheapest way of the UK cutting carbon emissions and bringing people out of fuel poverty. The government is accused of lacking ambition in this area: last parliament it prompted the insulation of 5 million homes. This parliament it aims to improve just 1 million.

The Green Deal reborn

In 2017 the Green Deal was relaunched — but this time provided by the private sector through a new Green Deal Finance Company. This scheme was replaced again in 2022 by the Green Homes Grant. This is a scheme for homeowners and landlords to support the installation of energy-efficient and low-carbon heating improvements to their homes. It was hoped this could help households save up to £600 a year on energy bills. In 2022 the need for less reliance on fossil fuels was emphasised due not only to the environmental impact, but also to the crisis in Ukraine and perceived over-reliance on Russian oil.

Follow-up questions

1. Explain why 'traditional economic policies', which create incentives by reducing the price of products, have had limited success.
2. Research what has happened to government energy policy in the UK since the announcement to scrap the Green Deal.
3. Find out how successful the Green Homes Grant has been.
4. Describe one other way in which households can be encouraged to conserve energy within their homes.

Application of Economics in the Real World 8.4 on the Green Deal describes a recent, but partially aborted, government approach to a possible environmental market failure. Contrast this with Application of Economics in the Real World 8.5, which describes a much older market failure that occurred over half a century ago in London.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.5

The great smog of 1952

Smog is a mixture of smoke and other pollutants, and of fog. Smog occurs when climatic conditions trap air above cities and industrial areas, which prevents the dirty air from dispersing. It was the Clean Air Act of 1956 which cleared most of the smoke out of the UK's city air. It is not surprising the Clean Air Act was passed, because the London smog of 1952 was the most deadly weather-related catastrophe in the UK for some 150 years. Figure 8.6 shows the amount of smoke emitted and the number of deaths in London between 1 and 15 December 1952.

Nevertheless, while sulphur dioxide pollution has decreased, at least in the lower atmosphere, and soot has largely disappeared, other less visible, but



At least 4,000 people are estimated to have died from the effects of the London smog

still toxic, chemicals are being emitted, particularly by motor vehicles which are powered by petrol and diesel fuel. Arguably, the most damaging pollutant is nitrous oxide, which is emitted unseen into the atmosphere by human activities such as agriculture, fossil fuel combustion, wastewater management and industrial processes. Nitrous oxide molecules stay in the atmosphere for an average of 120 years before disintegrating. The impact of 1 kilo of nitrous oxide on warming the atmosphere is over 300 times that of 1 kilo of carbon dioxide.

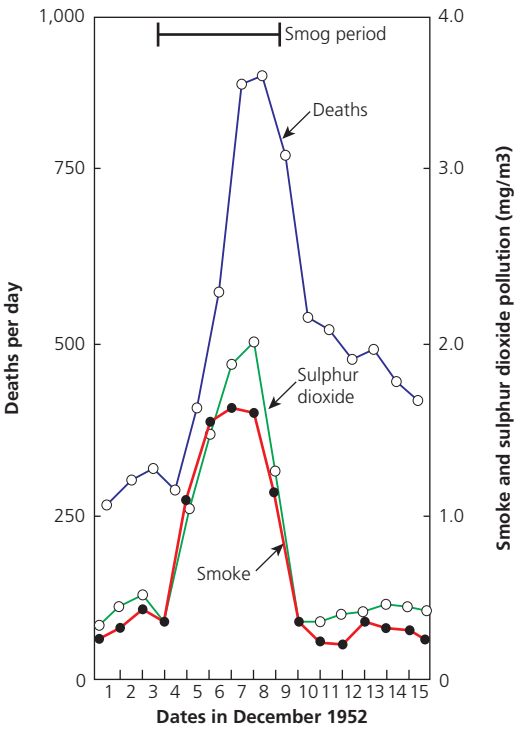


Figure 8.6 The effect of the London smog in 1952

There have been many other examples of serious air pollution in recent years. Rapid industrialisation and automobile ownership have led to serious smog problems in cities in China and India. In September 2015, the Volkswagen emissions scandal, so-called 'dieselgate', began. A US regulator found that Volkswagen had intentionally programmed diesel engines to activate their emissions controls only during laboratory emissions testing, but with up to 40 times more pollution being emitted in real-world driving conditions.

Follow-up questions

1. Analyse two ways in which a government can intervene to bring about cleaner air.
2. In recent years, many UK families have installed wood-burning stoves in their houses. Use the internet to find out whether the use of these stoves is causing atmospheric pollution in UK towns and cities.
3. Motor vehicle emissions break clean-air laws in the streets of many UK cities. Should all vehicles be banned from currently polluted streets? Justify your answer.
4. Find out what punishments have been levied on the Volkswagen car group in the USA and in other countries.

The concepts of marginal private, external and social costs and benefits

To explain the nature of externalities in greater depth, we need to apply the marginal analysis we used in the context of utility theory in Chapter 2 and in the explanation of marginal returns, cost and revenue in Chapter 4.

At the heart of microeconomic theory lies the assumption that, in a market situation, an economic agent considers only the private costs and benefits resulting from its market actions, ignoring any costs and benefits imposed on others. For the agent, private benefit maximisation occurs when:

marginal private benefit (*MPB*) = marginal private cost (*MPC*)

However, social benefit maximisation, which maximises the public interest or the welfare of the whole community, occurs when:

marginal social benefit (*MSB*) = marginal social cost (*MSC*)

Traditional economic theory usually assumes that households and firms seek to maximise their private benefit or self-interest, net of costs, and not the wider social interest of the whole community. They ignore the effects of their actions on other people. However, when externalities are generated, costs and benefits are inevitably imposed on others, so maximising net private benefit no longer coincides with the maximisation of net social benefit.

Social benefit is defined as private benefit plus external benefit. As a result:

marginal social benefit = marginal private benefit + marginal external benefit

(*MSB* = *MPB* + *MEB*)

Likewise, social cost is defined as private cost plus external cost, which means that:

marginal social cost = marginal private cost + marginal external cost

(*MSC* = *MPC* + *MEC*)

SYNOPTIC LINK

Refer back to Chapters 2 and 4 which first introduced the way marginal analysis is used in economic theory.

QUANTITATIVE SKILLS 8.1

Worked example: calculating social costs and benefits

Table 8.2 shows the costs and benefits of building a new shopping mall.

Table 8.2 Costs and benefits of building a shopping mall

Private cost	£10 million
External cost	£20 million
Private benefit	£8 million
External benefit	£25 million

Calculate the social cost and the social benefit and then decide whether building the shopping mall is worthwhile, first for society, and second for the property developer.

Social cost is calculated as follows:

social cost = private cost + external cost

and likewise social benefit is calculated as:

social benefit = private benefit + external benefit

Social cost is therefore £30 million, while social benefit is £33 million. Since the social benefit exceeds the social cost by £3 million, the shopping mall is worthwhile for society at large. However, it is not worthwhile for the property developer, whose private cost exceeds private benefit by £2 million.

TEST YOURSELF 8.5

Explain why B is the correct answer to the following question, and why A, C and D are incorrect.

Which one of the following creates an incentive for firms to reduce pollution?

A An increase in negative externalities when firms increase their output

B A tax levied on polluting firms

C Subsidies paid to firms that increase their levels of production

D The removal of regulations imposed on polluting firms

Using marginal analysis to show how negative production externalities cause market failure

We will now return to the example of pollution emitted by a fossil-fuel burning power station. We initially assumed that when a coal-burning power station generates electricity, only negative externalities are discharged and there are no positive externalities. Given this simplification, the marginal private benefit accruing to the power station from the production of electricity, and the marginal social benefit received by the whole community, are the same and shown by the downward-sloping curve in Figure 8.7. But, because pollution is discharged in the course of production, the marginal social cost of electricity production exceeds the marginal private cost incurred by the power station. In the diagram, the *MSC* curve is positioned above the *MPC* curve. The vertical distance between two curves shows the marginal external cost (*MEC*) at each level of electricity production.

The power station maximises private benefit by producing output Q_1, where *MPC* = *MPB*. Q_1 is immediately below point *A* in Figure 8.7. However, the socially optimal level of output is Q_2, where *MSC* = *MSB*. Q_2 is immediately below point *B* in Figure 8.7. The privately optimal level of output is thus greater than the socially optimal level of production. To put it another way, market forces over-produce electricity by the amount $Q_1 - Q_2$. The market fails because the power station produces the wrong quantity of the good, namely too much electricity. At the free-market price of P_1, electricity is too cheap. The price would have to rise to P_2 to bring about the socially optimal level of consumption.

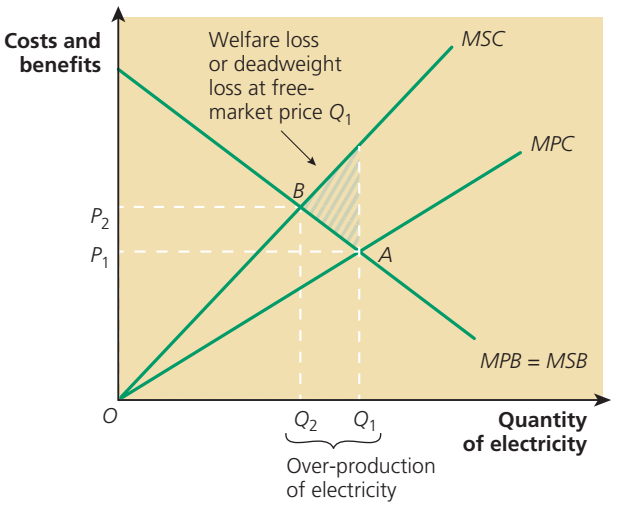


Figure 8.7 A coal-burning power station generating pollution (a negative production externality)

The shaded area in Figure 8.7 illustrates the 'loss' of welfare or deadweight loss (DWL), which exists at the free-market output, Q_1 (where *MPC* = *MPB*), all the way back to the socially optimum output, Q_2. This is because for units Q_2 to Q_1, the social cost of producing each unit of electricity is greater than the benefit society derives from each unit produced — that is, *MSC* > *MSB*. Society would be better off if these units, Q_2 to Q_1, were not produced and the resources were transferred to the production of other products. When production takes place at the socially optimal output (where *MSB* = *MSC*), the DWL is eliminated.

QUANTITATIVE SKILLS 8.2

Worked example: calculating private, social and external costs

Table 8.3 shows the costs incurred when producing different quantities of a chemical. Some of the cells in the table are blank.

Table 8.3 Costs incurred in chemical production

Output of chemicals (tonnes per week)	Marginal private cost (£000)	Marginal social cost (£000)	Marginal external costs (£000)
500	100		20
600		160	30
700	160	200	

Copy the table and insert the correct numbers in the blank spaces.

The completed version is shown in Table 8.4.

Table 8.4 Costs incurred in chemical production: completed version

Output of chemicals (tonnes per week)	Marginal private cost (£000)	Marginal social cost (£000)	Marginal external costs (£000)
500	100	**120**	20
600	**130**	160	30
700	160	200	**40**

The correct numbers to insert in the three blank spaces should be calculated by using the following equation:

MSC = MPC + MEC

In the top row: **£120,000** = £100,000 + £20,000

In the middle row: £160,000 = **£130,000** + £30,000

In the bottom row: £200,000 = £160,000 + **£40,000**

(Note: remember to include the thousands and the £ sign in your answers.)

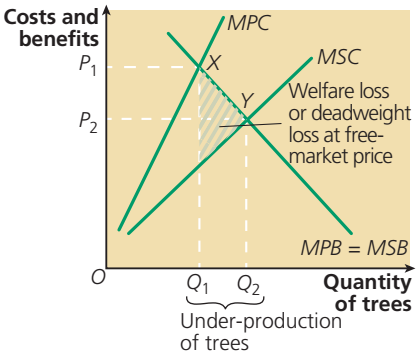


Figure 8.8 A commercial forestry company generating positive production externalities

Using marginal analysis to show how positive production externalities cause market failure

Whereas *negative* production externalities (illustrated in Figure 8.7) lead to the marginal social costs of production exceeding marginal private costs of production, when *positive* production externalities are generated, the marginal social costs of production lie *below* the marginal private costs incurred by the producers of the good or service. This is illustrated in Figure 8.8, which shows the costs incurred when a commercial forestry company plants trees.

The positive production externalities generated by tree planting include improved water retention in the soil and a carbon sink effect, whereby trees absorb greenhouse or global-warming gases from the atmosphere. As stated,

positive production externalities such as these mean that the *MSC* curve is positioned below the *MPC* curve in Figure 8.8. The vertical distance between the two curves shows a *negative* marginal external cost (*MEC*) at each level of tree planting. (A *negative* marginal *external cost* is exactly the same as a *positive* marginal *external benefit* enjoyed by society as a whole.)

In order to maximise its private benefit, the commercial forestry plants Q_1 trees. (This is the free-market level of output.) Q_1 is immediately below point *X*, where *MPC* = *MPB*. However, Q_1 is less than the socially optimal level of output Q_2, located below point *Y*, where *MSC* = *MSB*.

The shaded area in Figure 8.8 illustrates the 'loss' of welfare or deadweight loss (DWL), which exists at the free-market output, Q_1, all the way forward to the socially optimum output, Q_2. When production takes place at the socially optimal output, the DWL is eliminated. This is sometimes referred to as a 'welfare gain'. There is a net gain in social welfare if the trees between Q_1 and Q_2 are planted because the *MSB* is greater than the *MSC* for each of these trees up to the point at which *MSB* = *MSC*.

Figure 8.8 illustrates the fact that, when positive production externalities are generated, the market fails because too little of the good is produced and consumed. Under-production and under-consumption are depicted by the distance $Q_2 - Q_1$.

STUDY TIP

Note that in Figure 8.8 the 'spillover' benefits generated by production (positive external benefits) are depicted as a 'negative external cost'. The planting of trees reduces the costs imposed on society by pollution and hence the total cost to society of planting the trees is less than the cost to the firm. This is an important point to aid understanding of how to illustrate and interpret positive production externalities which will help you in the exam.

QUANTITATIVE SKILLS 8.3

Worked example: marginal costs and the benefits of vaccination

Table 8.5 shows some of the costs a consumer incurs and the benefits received by her family from receiving a vaccination.

Table 8.5 Costs and benefits of a vaccination

Marginal private cost of the vaccination (£)	Marginal private benefit resulting from the vaccination (£)	Marginal social benefit resulting from the vaccination (£)
20	100	1,000

(a) What is the marginal external benefit resulting from the vaccination?

Using the equation *MSB* = *MPB* + *MEB*, the marginal external benefit received by society from the vaccination is £900.

(b) Suppose evidence shows that a small proportion of people who are vaccinated suffer serious side-effects as a result of receiving the treatment. This creates a marginal social cost for society of £800 per vaccination. On the basis of these figures, should the health service recommend that people be vaccinated?

Given the extra information, the marginal social benefit of each vaccination exceeds the marginal social cost by £200 because marginal social benefit is £1,000 and marginal social cost is £800. On the basis of these figures alone, the health service should recommend that people are vaccinated.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.6

Road pricing

In recent years the issue of whether or not to charge motorists for the use of roads has entered public debate. The issue centres largely on road *congestion*, rather than on pollution, because fuel taxes are a better way of reducing the environmental pollution caused by vehicles burning fossil fuels. Since 2003 motorists have been charged for driving in central London during business hours, and a private-sector firm owns and runs a section of toll motorway north of Birmingham. Electronic pricing has become technically possible and is likely *eventually* to be used in future road-charging schemes. However, in the short term, the power of the motoring lobby, which argues that motorists should enjoy the freedom to use of the public highway without charge, is likely to delay the widespread introduction of road pricing, at least in the UK.

The case for and against road pricing brings together issues concerning both public goods and negative externalities. Roads are a good example of a quasi-public good, since toll booths or electronic pricing can be used to exclude free-riders. Road use also results in negative externalities. The extent to which negative externalities are produced depends in part upon whether the road is congested or uncongested. During peak periods, a rise in the number of road users increases the journey time of other road users. A negative externality is generated.

In the following analysis, we will assume that commercial vehicles, but not private cars, use the road. This means that all the negative externalities unleashed by road users are negative production externalities. However, road congestion could also legitimately be treated as a negative consumption externality — if generated by the use of private cars.

Given our simplifying assumption, Figure 8.9 illustrates the benefits and costs resulting from an extra commercial vehicle using the road, first when the road is uncongested, and then when road congestion has set in. The diagram shows the flow of traffic (e.g. the number of commercial vehicles travelling on the road per hour) on the horizontal axis and the cost per journey on the vertical axis.

When the traffic flow is less than F_1, an extra vehicle on the road does not impose any negative externalities or external costs upon other road users — provided we ignore the pollution emitted by the vehicles. In this situation, road use should be free, to encourage the allocatively efficient or socially optimal level of use. For levels of traffic flow between zero and F_1, the marginal social cost of road use equals the marginal private cost borne by vehicles and their owners ($MSC = MPC$). But once the road becomes congested (at flows of traffic greater than F_1), this is no longer the case, and there is a case for road pricing to provide the incentive to reduce road use.

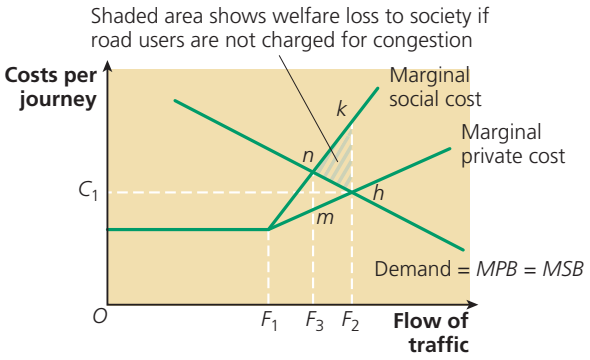


Figure 8.9 The benefits and costs of driving on an uncongested and a congested road

For traffic flows above F_1, each commercial vehicle driven on the road adds to traffic congestion, and all road users then suffer from increased journey times and the frustration of being stuck on a congested road. Beyond F_1, the marginal social cost of motoring is greater than the marginal private cost incurred by the last commercial vehicle to drive on the road ($MSC > MPC$). But in the absence of road pricing, when deciding whether or not to drive on the road, road users consider only the private cost of motoring and not the external cost dumped on others. If there is no charge for road use, commercial vehicles are driven up to traffic flow F_2 (at point h). At F_2, the marginal private benefit of driving the vehicle equals the marginal private cost ($MPB = MPC$). At this point, the private cost incurred by the marginal vehicle to use the road is C_1, but this is less than the social cost of the marginal journey, which includes the marginal cost of congestion caused by the marginal vehicle but suffered by other road users. At traffic flow F_2, the marginal external cost of congestion imposed on other road users is shown by the distance ($k - h$).

A misallocation of resources is the result. Commercial vehicles make more journeys than they would, had their owners or drivers borne the full social cost of using their vehicles. The shaded triangle bounded by the points *n*, *k* and *h* measures the welfare loss suffered by society at the privately optimal traffic flow F_2, where *MPB* = *MPC*.

Arguably, there is a case for road pricing when roads become congested. Allocative efficiency is improved when a motorist is charged a price equal to the marginal external cost imposed on other road users, as a result of the journey. The optimal congestion charge would be (*n* – *m*), which measures the marginal external cost of a journey at the socially optimal level of road use, F_3. Some congestion still occurs at F_3, but it is less than at F_2, and the owners or drivers of commercial vehicles pay for the congestion they generate. The congestion charge or road price internalises the externality. Journeys that are worth undertaking in the absence of a congestion charge are not worthwhile once the appropriate charge is imposed.

Follow-up questions

1 What is meant by the socially optimal level of road use?
2 Do you agree that fuel taxes are the best way of reducing the environmental pollution caused by vehicles burning fossil fuels? Justify your answer.
3 The passage states that electronic pricing has become technically possible. If this is the case, why have very few road-pricing schemes been introduced in the UK?
4 Suggest two other ways in which the environmental pollution caused by road traffic could be reduced.

Negative externalities and allocative efficiency

In Chapter 5, we explained how a perfectly competitive economy can, in theory at least, achieve a state of allocative efficiency when it brings about an outcome in which *P* = *MC* in all the markets that make up the economy. However, allocative efficiency could only occur if:

- there were competitive markets for all goods and services, including future markets
- there were no economies of scale
- markets were simultaneously in equilibrium.

However, it is impossible for markets throughout the economy — or indeed throughout the world economy — to meet all these conditions, so allocative efficiency is an abstract rather than a real-life concept. And in the context of market failure, we can now add a fourth requirement for allocative efficiency to be achieved:

- when *P* = *MC* there would have to be no externalities, negative or positive.

If we ignore the four bullet points set out above, in the long run, profit maximisation occurs in a perfect market at the price at which *P* = *MPC*. In the absence of externalities, this also means that the price equals the marginal social cost (*MSC*) of production: *P* = *MSC*. But as we have seen, when the production of a good causes pollution, external costs are generated, with the result that *MSC* > *MPC*. This means that when *P* = *MPC*, *P* < *MSC*.

To achieve allocative efficiency, price must equal the *true* marginal cost of production: that is, the marginal social cost and not just the marginal private cost. But in a market situation, profit-maximising firms are assumed only to take account of private costs and benefits. When externalities exist, therefore, the market mechanism fails to achieve an allocatively efficient outcome.

To put it another way, firms evade part of the true or real cost of production by dumping the externality on third parties. The price that the consumer pays for the good reflects only the private cost of production, and not the true cost, which

includes the external cost. The firm's output is thus under-priced, encouraging too much consumption. A misallocation of resources results: too much consumption, and hence too much production, means that too many scarce resources are being used by the industry that is producing the negative externalities.

STUDY TIP

While the specification states that you 'should be able to illustrate the misallocation of resources resulting from externalities in both production and consumption', the concept of allocative inefficiency is not mentioned in the context of market failures. You should nevertheless understand that resource misallocation means that there is an allocatively inefficient use of resources.

QUANTITATIVE SKILLS 8.4

Worked example: calculating an allocatively efficient price

The marginal private cost of producing an extra unit of a good is £15 and the marginal external cost resulting from the pollution incurred in the course of production is £3 for each extra unit.

Assuming no other externalities in consumption or production, all other things being equal, what price should achieve an allocatively efficient level of production and consumption of the good?

All other things being equal, the allocatively efficient level of production and consumption requires that price equals marginal social cost ($P = MSC$), with $MSC = MPC + MEC$. For this good, the MSC of producing an extra unit is £18 (£15 + £3), so a price of £18 is needed to ensure that the level of production and consumption of the good is allocatively efficient.

Note: in this example, since there are no externalities in consumption, the price of the product reflects the marginal social benefit (MSB) derived from the last unit of the product that is produced and consumed. Hence $P = MSC$ is equivalent to $MSB = MSC$.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.7

Allocative efficiency and rail fares

Motorists in cities pay substantially less than the costs they create when driving their vehicles. The greater the congestion, the truer this is. The extra cost or, as economists would call it, the marginal social cost of an extra vehicle coming onto a road is quantifiable. Wherever there is congestion, the marginal social cost will be greater than the marginal private cost, since the marginal costs to the motorist of using the road are additional vehicle costs, which include the cost of petrol or diesel fuel, and the costs of wear and tear and time. The motorist does not have to take into account the costs imposed on other road users and on pedestrians.

On the other hand, if rail transport in cities is required to cover its costs, it will be over-priced relative to users of urban roads, since rail users will be required to cover all the costs, both private and external, that they create, while road users will not. The effect of this difference in pricing policy is an inefficient distribution of traffic between road and rail. Less traffic travels by rail, especially at peak times, than is efficient. Similarly, there is too much traffic on the roads.

One way of getting prices right would be to raise the price of urban road use until both public and private road transport covered their true costs. But if we accept that it is politically imprudent or undesirable to raise the cost of using roads to a level where marginal social costs are covered, one can attempt to get the allocatively efficient relationship between road and rail by the opposite course of action: by keeping rail fares lower than they would be if the railways charged the full market price. This is the essence of the case for rail subsidies.

For this case to be proven, the traffic that the railways divert from the roads must reduce congestion by an amount sufficient to justify the rail subsidies required. Underlying this is the proposition that users of city roads pay less through taxation for using them than the full social cost of their use.

Follow-up questions

1 Define the term 'marginal social cost'.
2 The passage suggests that subsidising rail fares can establish the allocatively efficient relationship between the prices paid for road and rail use. Explain this statement.
3 Why does allocative inefficiency result in resource misallocation?
4 Outline the case against subsidising rail fares.

SECTION 8.4 SUMMARY

- Externalities exist when there is a divergence between private and social costs and benefits.
- The key feature of an externality is that there is no market in which it can be bought or sold.
- Externalities are produced and received outside the market, providing another example of a missing market.
- Externalities divide into production externalities and consumption externalities.
- A negative externality (external cost), such as pollution, is a public 'bad' dumped on others.
- A positive externality (external benefit), such as a beautiful view, is a public good that benefits others.
- Externalities lead to the 'wrong' quantity of a good being produced and consumed.
- If negative *production* externalities are generated, goods end up being too cheap or under-priced. Too much of the good ends up being produced and consumed.
- The opposite happens when firms generate positive production externalities. Prices end up being too high, and not enough of the good ends up being produced and consumed.
- In much the same way, when negative and positive *consumption* externalities are generated, markets are distorted and the wrong quantity of a good is produced and consumed.
- Environmental externalities often involve pollution of land, sea, rivers and air, and the externalities associated with road use and congestion.
- To explain externalities in greater depth, we need to apply marginal analysis.

8.5 Merit and demerit goods

Merit goods

Merit goods are not the same as public goods. As we have seen, public goods are defined by two characteristics: they are non-excludable and non-rival. Because of these characteristics, a market may fail to provide a pure public good such as national defence. By contrast, markets can and do provide merit goods such as healthcare and education but, arguably, they under-provide them. While public goods can result in a missing market or *complete* market failure, merit goods (and also their opposite, demerit goods) lead to *partial* market failure. As goods such as private healthcare and private education clearly show, markets provide merit goods, but they provide the 'wrong' quantity.

KEY TERM

merit good a good, such as healthcare, for which the social benefits of consumption exceed the private benefits. Value judgements are involved in deciding that a good is a merit good.

STUDY TIP

Students often assert that any good that is 'good for you' is a merit good. Make sure you can explain why this assertion is wrong.

KEY TERMS

information failure occurs when people make wrong decisions because they do not possess or they ignore relevant information. Very often they think just of the present rather than the future consequences of their actions.

social benefit the total benefit of an activity, including the external benefit as well as the private benefit. Expressed as an equation: social benefit = private benefit + external benefit.

A merit good has two important characteristics: positive externalities in consumption and **information failures** which distort a consumer's choice of what is the privately optimal level of consumption.

- **Merit goods and positive externalities.** When a person consumes a merit good such as cycling to work, the resulting positive externalities benefit other people. The additional benefit to society is less traffic and less air pollution. The **social benefit** enjoyed by the wider community is greater than the private benefit enjoyed by the healthy cyclist.
- **Merit goods and information failures.** For a merit good such as healthcare, the long-term private benefit of consumption exceeds the short-term private benefit of consumption. But when deciding how much to consume, individuals take account of short-term costs and benefits, ignoring or under-valuing the long-term private costs and benefits. For example, at the start of a student's A-level course, the teachers may advise their students to study little and often from the very start of their courses in order to embed their learning over a longer period of time. First-year students may not think that they need to do this and, as a result, will find that there is an overwhelming amount to learn in their second year in the run-up to their final exams.

Education and healthcare are the best-known examples of merit goods. However, many goods can be classified as merit goods, though you must avoid the temptation that many students succumb to, to define any good that is 'good for you' as a merit good. Most consumer goods are good for you, but economists don't classify them as merit goods. Besides education and healthcare, other examples of merit goods, most people would argue, are car seatbelts, crash helmets, public parks and museums. Nevertheless, the classification of merit and demerit goods depends upon value judgements — what one person regards as a merit good, for example opera, may be viewed by another person as a demerit good.

TEST YOURSELF 8.6

Explain why C provides the correct answer to the following question, and why A, B and D are incorrect.

A difference between a merit good and a public good is that:

A a merit good is always provided by the private sector whereas a public good is always provided by the government

B a merit good has a cost attached to it, whereas a public good is provided without cost to anybody

C consumption of a merit good reduces the amount available for others whereas an individual's consumption of a pure public good leaves unaffected the amount available for others

D a merit good is limited in supply whereas the supply of public goods is infinite

STUDY TIP

You should understand and be able to explain that merit goods are private goods; they are excludable and rival, even though they are often provided by the public sector.

Positive consumption externalities and merit goods

A merit good, such as education or healthcare, is a good or service for which the social benefits of consumption enjoyed by the whole community exceed the private benefits received by the consumer. Consumption by an individual arguably produces positive externalities that benefit the wider community. The community benefits from an educated population, and a healthy population means there are fewer people to catch diseases from.

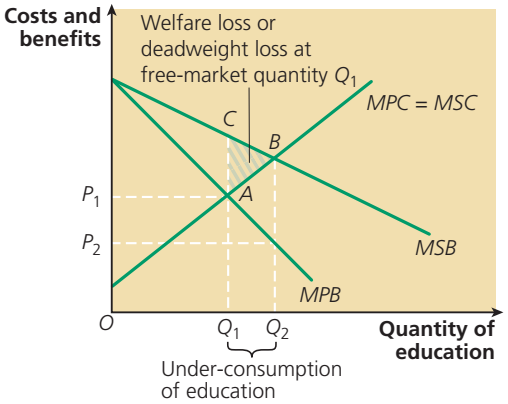


Figure 8.10 Under-consumption of a merit good in a free market

If educational services were to be provided solely through the market, and at market prices, too few people would benefit from education. In Figure 8.10, the privately optimal level of consumption is Q_1, determined at point *A*, where *MPC* = *MPB*. This is below the socially optimal level of consumption, Q_2. The socially optimal level of education lies below point *B* on the diagram, located where *MSC* = *MSB*. According to this analysis, free-market provision of merit goods leads to under-consumption. The *MSB* of units of education between Q_1 and Q_2 is greater than the *MSC* of these units. Producing and consuming these units will add to the welfare of society. Suppose that the government reacts by providing a **subsidy** which reduces the price of education to P_2. At the subsidised price, consumption of education rises to the socially optimal level of Q_2. The market failure has now been corrected.

KEY TERM

subsidy a payment made by government or other authority, usually to producers, for each unit of the subsidised good that they produce. Consumers can also be subsidised: for example, bus passes given to children to enable them to travel on buses free or at a reduced price.

STUDY TIP

Make sure you practise drawing graphs such as the ones in Figures 8.7, 8.8, 8.10 and 8.11. These are fundamental to analysing externalities in production and consumption and are key diagrams for the A-level course.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.8

Museums as merit goods

Museums perform the important cultural function of conserving, interpreting, researching and displaying heritage. Museums have a mix of ownership patterns. For example, over 40% of UK museums are governed by public authorities, with the rest privately owned, mostly on a non-profit basis. Museums cover a wide range of institutions of varying size and reputation, ranging from internationally renowned institutions such as the British Museum to a large number of relatively small, often locally focused museums.

The funding of museums remains a source of considerable debate. Government subsidy or provision may be justified on the basis that museums generate external benefits: for example, knowledge acquired by a visitor may be passed on to others. Museums may also be regarded as merit goods, generating a better educated and informed public and collective pride.



Museums may be regarded as merit goods that should be subsidised

Government subsidy of museums may also be justified on the grounds that the subsidy would encourage consumption. The case against government subsidy is that it may encourage inefficiency, leading to government failure through favouring well-off visitors who can afford to pay. It is also a cost to the taxpayer.

Follow-up questions

1. Do you agree that a museum is a merit good? Justify your answer.
2. Some argue that if entry to museums is free, people will not value what museums have to offer. Explain why you agree or disagree.
3. 'Some people believe that when free, museums attract the wrong sort of visitor, namely homeless people looking for warmth. Entry prices should be used as a rationing device so that only people who genuinely want to see what a museum has to offer are admitted.' Should prices be used as a rationing device in this way?
4. In the USA, many museums and art galleries, such as the Getty Museum in Los Angeles, are funded by rich philanthropists. Should British museums be funded by rich individuals rather than by the state?

Demerit goods

As their name suggests, **demerit goods** are the opposite of merit goods. The **social costs** to the whole community which result from the consumption of a demerit good, such as tobacco or alcohol, exceed the private costs incurred by the consumer. This is because consumption by an individual produces negative externalities that harm the wider community. The *private* cost can be measured by the money cost of purchasing the good, together with any health damage suffered by the person consuming the good. But the *social* costs of consumption also include the cost of the negative externalities. These include, for example, costs imposed on other people from passive smoking and road accidents caused by drunken drivers, together with the cost of taxes raised to pay for the care of victims of tobacco- and alcohol-related diseases.

As is the case with merit goods, there are two main characteristics of a demerit good, the first centring on externalities, and the second on information failures that affect the good's consumption. With the first characteristic, the benefit the smoker or drinker derives from consumption is greater than the benefit

KEY TERMS

demerit goods goods, such as tobacco, for which the social costs of consumption exceed the private costs. Value judgements are involved in deciding that a good is a demerit good.

social cost the total cost of an activity, comprising the external cost as well as the private cost. Expressed as an equation: social cost = private cost + external cost.

to society. The second characteristic is based on the distinction between the short-term and long-term net benefit derived by the person consuming a demerit good.

- **Demerit goods and negative externalities.** When a person consumes a demerit good such as tobacco, negative externalities are generated which are unpleasant or harmful for other people. People unwillingly breathe in the fumes the smoker discharges, with eventual harmful effects on their health. (This is the problem of passive smoking.) Smelly clothing is a more trivial example of a negative externality in consumption that is caused by smoking. Hence, the marginal social benefit is less than the marginal private benefit.
- **Demerit goods and information failures.** When teenagers first get the 'habit' of smoking, drinking or drug taking, they may either ignore the long-term private costs they may suffer later in life, or downplay the significance of these costs. Either way, young people tend to be short-sighted with respect to the costs of consuming demerit goods. A person who started drinking as a teenager may regret the decision later in life when suffering an alcohol-related illness. Hence, the long-run net benefit of consuming a demerit good is less than the short-run benefit.

Merit and demerit goods and value judgements

The left-hand and right-hand columns of Table 8.6 list a number of goods that are accepted by most people as clear-cut examples of merit goods or demerit goods. However, for the goods listed in the middle column — for example, contraception — the position is less clear. Because people have different values and ethics, contraception and abortion are viewed by some people as merit goods, but by others as demerit goods. Whether a good is classified as a merit good or a demerit good, or indeed as neither, thus depends on the value judgements of the person making the classification.

Table 8.6 Merit and demerit goods, and less clear-cut cases

Merit goods	Merit or demerit goods?	Demerit goods
• Education • Healthcare (e.g. vaccination, preventative dental care, HIV testing) • Crash helmets • Car seatbelts • Museums and public parks	• Contraception • Abortion • Sterilisation	• Tobacco • Alcohol • Narcotic drugs (e.g. heroin, crack cocaine) • Pornography • Prostitution

It is also worth noting that under-provision of merit goods and over-provision of demerit goods may also result from the imperfect information that consumers possess, which can lead — for example, in the case of demerit goods — to failure to give adequate weight to the long-term consequences of consuming a good such as tobacco. We will explore the effects of imperfect information a little later in the chapter.

SYNOPTIC LINK

The importance of value judgements when distinguishing between merit and demerit good provides an important example of the distinction between positive statements and normative statements, first explained in Chapter 1. To recap, a positive statement is a statement of fact or a statement that can be tested to see if it is right or wrong. For example, statements that the Earth is round and that the Earth is flat are both positive statements, though scientific evidence shows that the second statement is wrong. By contrast, a normative statement is a statement of opinion, involving a value judgement. Thus, a statement that the Earth *ought* to be flat is normative.

Are there such things as demerit goods and merit goods?

Some economists, usually of a pro-free-market persuasion, question whether any goods are demerit goods or merit goods. These economists are libertarians (those who believe that government intervention should be minimal) who believe that individuals are the best judges of what is bad or good for them. People should be free, within the law, to smoke or drink themselves to death, or to choose whether to have a vaccination to protect them against the potentially life-threatening effects of Covid-19. According to this view, it is not the state's role to intervene in individual behaviour, except when the pursuit of self-interest harms other people. For example, the state should prevent motorists from exercising the freedom to drive on the wrong side of the road.

Anti-libertarians, however, do not accept this view of individual freedom of choice. They argue that over-consumption of demerit goods and under-consumption of merit goods invariably leads to other people being harmed and that the libertarian argument is misleading. According to this argument, taxation, subsidy and regulation are fully justified as means of altering individual behaviour.

Lying somewhat in between these two approaches are the insights of behavioural economists, who argue that individuals should be given as much choice as possible in deciding what is good for them, but that governments should use appropriate policies to 'nudge' them in the direction deemed socially optimal.

Using marginal analysis to show how negative consumption externalities cause over-consumption of demerit goods

In the same way as the consumption of merit goods generates positive externalities which benefit the wider community, the consumption of demerit goods leads to the dumping of negative externalities on others.

According to the first characteristic of a demerit good given on page 255, the consumption of goods such as tobacco and alcohol by an individual leads to the dumping of negative externalities on other individuals or third parties. Figure 8.11 shows that too much tobacco is consumed when bought at market prices unadjusted by taxes or by a minimum price law. At least in the short term, the privately optimal level of consumption is Q_1, where $MPC = MPB$.

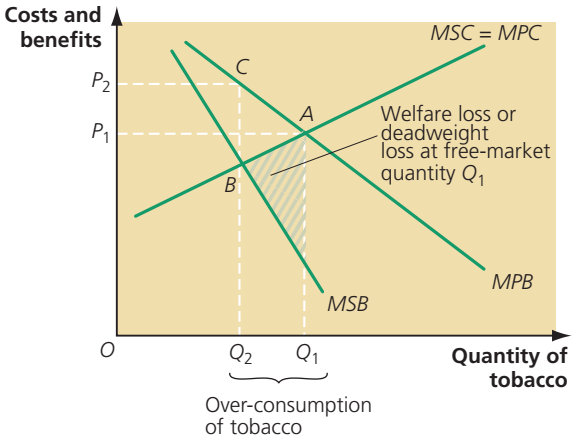


Figure 8.11 How over-consumption of a demerit good occurs in a free market

This is greater than the socially optimal level of consumption, Q_2, located where *MSC* = *MSB*. Free-market provision of demerit goods therefore leads to over-consumption, and hence over-production of tobacco.

The negative externalities in this example are generated when people smoke (e.g. the smoke breathed in by passive smokers who do not enjoy the fumes they inhale), so the externalities are *consumption* externalities rather than *production* externalities. As a result, the marginal social benefit (*MSB*) curve of the whole community lies below the marginal private benefit (*MPB*) curve of the smokers themselves, with the distance between the two curves showing the negative externality.

To bring about the socially optimal level of consumption at Q_2, an indirect tax equal to the distance between points *B* and *C* in Figure 8.11 could be imposed, which would increase the price of tobacco to P_2. The tax needs to be set equal to the marginal external cost (*MEC*) of consumption at the socially optimal level of consumption (Q_2).

STUDY TIP

It is important to understand that the costs imposed by consumers on society (the negative externalities) are depicted in Figure 8.11 as negative benefits. Thus, the distance between points *B* and *C* on the graph can be interpreted either as the marginal external cost of smoking or as the negative external benefit of smoking at the socially optimal level of smoking, Q_2. The two are the same. For example, if someone is smoking at a party, the pleasure derived from the event by non-smokers is reduced; hence the cumulative benefit to all the party-goers is less than the pleasure derived by the smoker.

Note also that, for the socially optimal level of smoking to be zero (in which case a ban might be imposed on smoking), the *MSB* curve in Figure 8.11 would need to positioned well to the left of the curve in the diagram, intersecting the *MSC* curve at zero quantity.

TEST YOURSELF 8.7

Table 8.7 shows the marginal private and external benefits and the marginal private and external costs of two products at their free-market equilibrium levels of output.

Table 8.7 Private and external costs and benefits of products W and Z

	Product W (£)	Product Z (£)
Marginal private benefit	400	140
Marginal external benefit	100	80
Marginal private cost	400	140
Marginal external cost	180	60

Taking into account the information provided in the table, for each of the products W and Z, is there a case for increasing the level of output above the free-market output, or is there a case for reducing the level of output so that it is below the free-market output?

SYNOPTIC LINK

Taxes and government spending are used in the government's fiscal policy. Along with monetary policy, fiscal policy is usually considered to be a macroeconomic topic. As such, fiscal policy is explained in some depth in Chapter 13. Nevertheless, taxes imposed on firms and subsidies, which are part of government spending, also have microeconomic dimensions.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.9

Smoking yourself 'fit'

Strange as it seems, early cigarette advertisements, such as the Kensitas advert from 1929 (see the photo on the left), often boasted the 'health benefits' of smoking, claiming 'relief' from asthma, wheezing, hay fever and obesity. In 1946 the US tobacco company Camel ran a series of adverts claiming that Camel cigarettes were the 'doctor's choice' (see the photo on the right).

By the 1950s, research began to link smoking to cancer. Worldwide, tobacco use causes more than 7 million deaths per year, and current trends show that tobacco use will cause more than 8 million deaths annually by 2030, largely through the growth of smoking in developing countries. On average, smokers die ten years earlier than non-smokers.

Cigarettes (and other tobacco products) and alcoholic drink are the two best-known examples of demerit goods. These days, governments in countries such as the UK either ban tobacco and drinks advertising, or severely regulate what the adverts can show. This has not always been the case, as the photographs show.





Magazine advertisements for Kensitas, 1929 (left) and for Camels, 1946 (right) claim the health benefits of smoking

Tobacco and drinks companies want to make their products appealing to young people, possibly in the hope that, if teenagers develop the habit, they will be hooked on cigarettes and alcohol for the rest of their lives. To combat claims that the industry has been acting in an irresponsible way, the UK firms that produce alcoholic drinks have set up a public relations organisation, the Portman Group, to monitor adverts that might bring the drinks industry into disrepute.

Follow-up questions

1 Most economists agree that tobacco is a demerit good. Why is this so?
2 Why do governments in countries such as the UK and USA now ban advertisements like these?
3 Is self-regulation by trade bodies such as the Portman Group the best way of discouraging consumption of demerit goods? Justify your answer.
4 Find out how successful e-cigarettes have been in reducing consumption of nicotine. Are there chemicals in e-cigarettes which may be harmful?

Merit and demerit goods and information failure

Some economists argue that under-consumption of merit goods and over-consumption of demerit goods stem not so much from the externalities that consumption generates, but from an information failure. Many people become addicted to demerit goods in their teenage years. Because of peer-group pressure and related factors, teenagers are heavily influenced by factors relating to lifestyle and personal circumstances, while at the same time ignoring, downplaying or being short-sighted about how their addictions may affect them many years ahead. Individuals take account of *short-term* costs and benefits, but ignore or under-value the *long-term* private costs and benefits.

For most merit goods, the long-term private benefit of consumption exceeds the short-term private benefit. But, as is the case with demerit goods, when deciding how much to consume, individuals often take account only of short-term costs and benefits. Preventative dentistry provides a good example. Many people ignore the long-term benefit of dental check-ups, and decide not to consume the service. Their decisions are influenced by factors such as the price charged by the dentist and the unpleasantness of the dental experience. Unfortunately, they may end up later in life with rotten teeth or gum disease, saying: 'If only I had visited the dentist more often when I was younger.'

SYNOPTIC LINK

Section 2.2 of Chapter 2 discusses the importance of information for decision making and the effect of asymmetrical information. Information imperfections such as asymmetric information can be a key source of market failure.

EXTENSION MATERIAL

Merit goods and uncertainty, moral hazard and adverse selection

Uncertainty about future long-term benefits and costs contributes to under-consumption of merit goods. For example, a person does not know in advance when, if ever, the services of a specialist surgeon might be needed. Sudden illness may lead to a situation in which a person cannot afford to pay for costly surgery, if provided solely through a conventional market.

One market-orientated solution is for private medical insurance to pay for the cost of treatment at the time when it is needed. However, private medical insurance often fails to pay for treatment for the chronically ill or for the poor. Private insurance may also fail to provide medical care for risk-takers in society who decide not to buy insurance, as distinct from risk-averters, who are always the most-ready customers for insurance.

Like all private insurance schemes, healthcare insurance suffers from two further problems, both of which lead to market failure. These are the problems of **moral hazard** and adverse selection. Moral hazard is demonstrated by the tendency of people covered by health insurance to be less careful about their health because they know that, in the event of accident or illness, the insurance company will pick up the bill. Adverse selection relates to the fact that people whose health risks are greatest are also the people most likely to try to buy insurance policies. Insurance companies react by refusing to sell health policies to those who most need private health insurance. For those to whom they do sell policies, premium levels are set sufficiently high to enable the companies to remain profitable when settling the claims of customers facing moral hazard or who have been adversely selected.

Public collective provision, perhaps organised by private-sector companies but guaranteed by the state and funded by compulsory insurance, may be a better solution. Both private and public collective-provision schemes are a response to the fact that the demand or need for medical care is much more predictable for a large group of people than for an individual.

KEY TERM

moral hazard the tendency of individuals and firms, once insured against some contingency, to behave so as to make that contingency more likely.

SYNOPTIC LINK

Moral hazard is mentioned in the specification in the context of the regulation of the financial system and monetary policy (see Chapter 12). Make sure you understand the concept both in the context described in this chapter, of persons covered by health insurance failing to look after their health, and also the context described in Chapter 12, of banks making risky investments because they know that, if an investment goes wrong, they will be bailed out by the government and/or the central bank. This means that banks will be saved from failure by using government or public funds.

TEST YOURSELF 8.8

With the use of examples, explain the difference between a merit good and a public good.

SECTION 8.5 SUMMARY

- Markets provide merit goods such as healthcare and education but generally they under-provide.
- When a person consumes a merit good, the resulting positive externalities benefit other people.
- The long-term private benefit resulting from consuming a merit good exceeds the short-term private benefit of consumption.
- When a person consumes a demerit good, the resulting positive externalities harm other people.
- The social costs to the whole community which result from the consumption of a demerit good, such as tobacco or alcohol, exceed the private costs incurred by the consumer.
- Some economists believe that under-consumption of merit goods and over-consumption of demerit goods stem not so much from the externalities that consumption generates, but from an information failure.

8.6 Market imperfections

The effect of imperfect and asymmetric information

Whereas all consumers and firms would possess perfect information about what is happening in a perfectly competitive market, the same is not true in imperfect competition. For example, in the context of an oligopolist's kinked demand curve, we explained in Chapter 5 how firms in imperfectly competitive markets are likely to lack accurate information about how their rivals, namely other firms in the market, may react to their pricing decisions. To get rid of the resulting uncertainty, imperfectly competitive firms may be tempted to collude (cooperate unlawfully). If cartels are formed, this can lead to an outcome which displays the disadvantages but not the advantages of monopoly. As the next section explains, by restricting output and raising prices, monopoly is a form of market failure.

In Chapter 3, Application of Economics in the Real World 3.6 on the market for second-hand cars illustrated the effects of asymmetric information on market outcomes. The sellers of second-hand cars possess more information about the quality of the cars than prospective buyers. By driving down the prices of good-quality cars below what they would otherwise be, the market is distorted, which results in a form of market failure.

SYNOPTIC LINK

Chapters 3 and 5 provide examples of the effects of asymmetric information and imperfect information. Reading these chapters will give you a deeper understanding of the meaning and implications of imperfect information and how it can lead to various forms of market failure.

Why the existence of monopoly may lead to market failure

We also saw in Chapter 5 how, compared to competitive markets and in pursuit of excess profit, monopolies are likely to restrict market output and raise prices. The result is consumer exploitation. Prices end up being too high, with too few of society's scarce resources being allocated to the market in which the monopoly is producing. However, as Chapter 5 also explains, sometimes the benefits of monopoly can exceed the costs, in which case there is no market failure.

SYNOPTIC LINK

Refer back to the discussion of the advantages and disadvantages of monopoly in Chapter 5, section 5.5. This will help you to take a theoretical view as to whether the potential for market failure caused by monopolies outweighs the possible benefits to society.

STUDY TIP

Make sure you can link the analysis of market failure to the reasons why monopoly and market imperfections provide examples of market failure: for example, the statement in Chapter 5 that the existence of deadweight loss (DWL) provides evidence of market failure in monopoly.

TEST YOURSELF 8.9

With the help of a diagram, explain why a monopoly restricts output and raises its selling price above the price that would be charged by a firm in a perfectly competitive market.

Why the immobility of factors of production can lead to market failure

KEY TERM

immobility of labour the inability of labour to move from one job to another, either for occupational reasons (e.g. the need for training) or for geographical reasons (e.g. the cost of moving to another part of the country).

An important cause of market failure is the immobility of factors of production, which can involve either geographical or occupational immobility. We summarise below the issues of **immobility of labour**.

We explained in Chapter 6, on page 181, some of the causes of the occupational and geographical immobility of labour. As we stated there, occupational immobility of labour arises when workers lack the skills to move between different types of employment, and because expensive and time-consuming training may be necessary if workers are to switch. The need to gain recognised professional qualifications acts as an important cause of occupational immobility, especially when workers want to move between different countries. To protect their own workers, governments often fail to recognise other countries' professional qualifications.

TEST YOURSELF 8.10
Give one example of the occupational immobility of labour and one example of the geographical immobility of labour.

We also stated that geographical immobility of labour occurs when geographical barriers between markets prevent labour moving from one area to another to find employment. Especially important are immigration laws which act as a barrier to the international mobility of labour between countries.

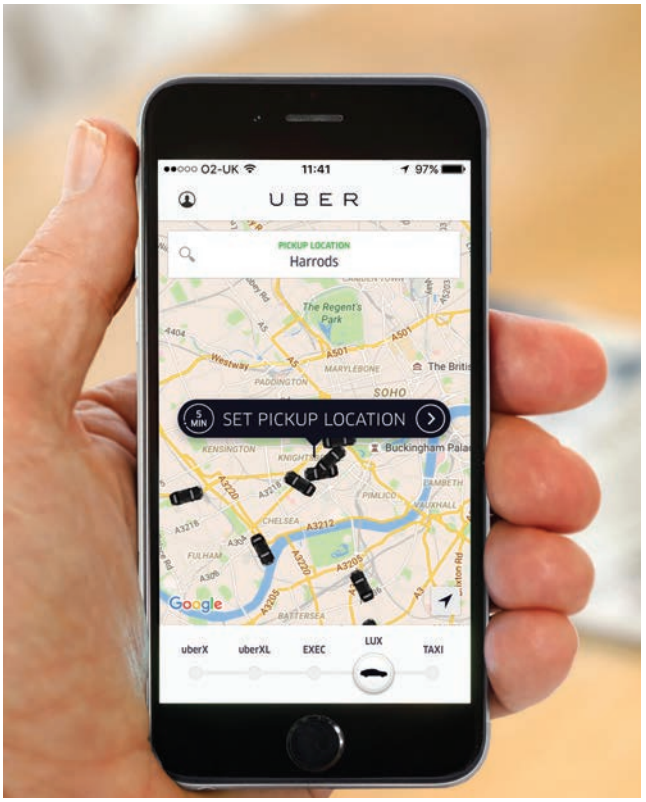
All these causes of immobility of labour lead to unemployment and a waste of scarce resources, and contribute to market failure in factor markets.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.10

Taxi wars

Until quite recently, the London taxi market, in which cabs could be hailed or flagged down in the street, was dominated by black cabs. Entry into the black cab driving profession was strictly limited by the requirements, first, to spend several years gaining 'the knowledge' of all the streets in the centre of the country's capital city, and second, to be granted a cab-driver's licence, which were deliberately kept in short supply.

Competition first emerged from mini-cabs, but they had to be pre-booked by phoning or visiting a mini-cab company's office, and then hoping that the car would show up. But a new form of competition has emerged which threatens to destroy the black cab market. Passengers are no longer subject to the availability of a black cab or a pre-booked mini-cab. They can now make use of a smartphone app that essentially puts a taxi in their pocket whenever they need one, in any major city in the world.



Uber has revolutionised employment in the London taxi market

By reducing barriers to labour market entry, smartphone technology is destroying the monopoly power of black cab drivers. Since the launch of the Uber app in 2009, there has been a massive power shift in the cab market. Uber and other web-based cab-booking services are accessed via smartphones. The service connects passengers with vetted, private drivers who pick up their customers within minutes and take them to where they want to go. When the booking is made, the passenger's location is pinpointed with the GPS on their phone. The passenger chooses the kind of car they need, checks the estimated price of the journey, and is then told how long they will need to wait for their car to arrive. They see a photo of the driver, their name and a contact phone number. When arriving at their destination, passengers do not need to hand over cash, as payment automatically takes place via a credit card. Next day an email is sent to the passenger with a breakdown of the costs, and passengers are not expected to tip their driver.

Follow-up questions

1 Before Uber entered the market, in what ways did black cab companies possess monopoly power?
2 Explain the effect of the Uber app on labour mobility in the cab trade.
3 In 2017, Transport for London (TfL) refused to renew Uber's licence to operate in London. Why did TfL do this, and what has happened since?
4 Describe one other example of smartphone technology improving labour market mobility.

SECTION 8.6 SUMMARY

- Consumers and firms do not possess perfect information about what is happening in an imperfectly competitive market.
- Firms in imperfectly competitive markets are likely to lack accurate information about how their rivals, namely other firms in the market, may react to their pricing decisions.
- To get rid of the resulting uncertainty, imperfectly competitive firms may be tempted to collude, and in some cases to form cartels.
- A cartel has the disadvantages but not the advantages of monopoly. As the next section explains, by restricting output and raising prices, monopoly is a form of market failure.
- Immobility of labour is the inability of labour to move from one job to another, either for occupational reasons (e.g. the need for training) or for geographical reasons (e.g. the cost of moving to another part of the country).

8.7 Competition policy

The theoretical background to competition policy

Two important points to remember about perfect competition are:

- In the absence of economies of scale, perfect competition would be more productively and allocatively efficient than monopoly.
- In perfect competition, consumer sovereignty rules, whereas monopoly leads to the manipulation of consumers and the exploitation of producer sovereignty. By restricting output and raising prices, monopoly results in a net welfare loss as well as a transfer of consumer surplus into producer surplus and monopoly profit.

But even though perfect competition does not actually exist in real-world economies, when firms operate in markets where a high degree of competition prevails, many economists believe that this helps to prevent the abuses that can occur when firms possess a significant degree of monopoly power.

KEY TERM

competition policy the part of the government's microeconomic policy and industrial policy which aims to make goods markets more competitive. It comprises policy towards monopoly, mergers and restrictive trading practices.

As its name implies, **competition policy** is the part of government economic policy which tries to make the imperfectly competitive and monopolistic markets of the real world more competitive. The aims of competition policy include preventing the exploitation of monopoly power, reducing costs of production, improving efficiency, getting rid of excessive profit so that prices reflect costs of production, and removing entry and exit barriers which separate markets.

Competition policy does recognise, however, that there are two main circumstances in which monopoly may be preferable to small firms producing in a competitive market:

- When the size of the market is limited but economies of scale are possible, monopolies can produce at a lower average cost than smaller, more competitive firms.
- Under certain circumstances, firms with monopoly power may be more innovative than firms that are not protected by entry barriers. When this is the case, monopoly may be more dynamically efficient than a more competitive market.

The three different parts of UK competition policy

As noted in the key term definition above, competition policy in the UK comprises policy towards monopoly, mergers and restrictive trading practices.

Monopoly policy

KEY TERM
Competition and Markets Authority (CMA) government agency responsible for advising on and implementing UK competition policy.

Since there are few, if any, pure monopolies in the UK, monopoly policy might better be called 'oligopoly policy' or policy towards highly concentrated markets dominated by just a few firms. Of the three main elements of competition policy we examine in this chapter, monopoly policy is the one with the longest history, dating back well over half a century to the establishment of the Monopolies Commission in 1948. The Monopolies Commission later became the Competition Commission, which in turn in 2013 was rebranded as the **Competition and Markets Authority** (CMA). The CMA began operations in 2014. The CMA is not restricted solely to the investigation of pure monopoly. It also investigates mergers that might create a new monopoly or a highly concentrated market structure, which is where one firm, or a small number of firms, have a high share of the market.

STUDY TIP
The label 'monopoly policy' is slightly misleading as very often the policy is aimed at oligopolies or concentrated markets rather than at pure monopoly.

A cost–benefit approach to monopoly policy

TEST YOURSELF 8.11
What is meant by a 'cost–benefit' approach to the problem of monopoly?

Economists recognise that monopoly can be either good or bad depending upon circumstances, so UK policy has always been based on the view that each case must be judged on its merits. If the likely costs resulting from the reduction of competition exceed the benefits, monopoly should be prevented, but if the likely benefits exceed the costs, monopoly should be permitted. Ongoing regulation is needed to make sure that firms, particularly large firms, continue to act in the public interest.

The Competition and Markets Authority (CMA)

As already noted, since 2014 UK monopoly policy has been implemented by the Competition and Markets Authority (CMA). The CMA was formed through the merger of two older government agencies, the Office of Fair Trading (OFT) and the Competition Commission (CC). Since 2017, the CMA has been responsible to a government ministry, the Department for Business, Energy and Industrial Strategy (BEIS).

The CMA uses market structure, conduct and performance indicators to scan or screen the UK economy on a systematic basis for evidence of monopoly abuse. Concentration ratios provide evidence of monopolistic market structures. Market conduct indicators such as consumer and trade complaints allow the CMA to monitor anti-competitive business behaviour. When the CMA discovers evidence of monopoly that it believes is likely to be against the public interest, it investigates further. Until recently, the main issue to be

decided was the relatively narrow one of whether particular trading practices undertaken by the investigated firm(s) were in the public interest. The 'public interest' was fairly vaguely defined. The 2002 Enterprise Act changed this, introducing competition-based tests to replace the public interest test. The tests centre on whether any features of the market (which include structural features and conduct by firms or customers in the market) prevent, restrict or distort competition.

Alternative approaches to the problem of monopoly

Ever since the start of competition policy over 70 years ago, the UK has adopted a regulatory and investigatory approach to the problem of monopoly. Relatively few firms and takeover bids are actually investigated. The policy rationale is that the possibility of a CMA investigation creates sufficient incentive for most large firms to behave well and to resist the temptation to exploit monopoly power in ways that are against the public interest.

However, although the CMA has adopted a 'watchdog investigatory/regulatory' role, a number of other strategic approaches could, in principle, be used to deal with the problem of monopoly. These include:

- **Compulsory breaking up of all monopolies ('monopoly busting').** Many free-market economists believe that the advantages of competitive and free markets, namely economic efficiency and consumer sovereignty, can be achieved only when the economy is as close as possible to perfect competition. In itself, they argue, monopoly is bad and impossible to justify. The government should therefore adopt an automatic policy rule to break up monopolies wherever they are found to exist. UK policy-makers have rarely, if ever, adopted a monopoly-busting approach, although, as we have explained, powers do exist that allow the government to order the break-up of an established monopoly.
- **Use of price controls to restrict monopoly abuse.** Although price controls have been used by UK governments at various times to restrict the freedom of UK firms to set their own prices, this policy has only been used in a limited way in recent years. Under the influence of free-market economic theory, price controls have fallen out of favour.
- **Taxing monopoly profits.** As well as controlling prices directly, the government can tax monopoly profit to punish firms for exploiting their monopoly power and making excessive profit. Monopoly taxes have not generally been used in the UK, except on a few occasions — for example, on the 'windfall' gain that landlords receive when the land they own is made available for property development. Similarly, windfall profits received by banks from high interest rates have been subject to a special tax.

STUDY TIP

You don't need to learn in detail all the policies listed here. Knowledge of two or three should do. However, try to think of relevant diagrams to illustrate the ones you choose.

- **Rate of return regulation.** In the USA, the regulators have imposed maximum rates of return on the capital that the utility companies employ. In principle, these act as a price cap, as the utilities are fined if they set prices too high and earn excessive rates of return. However, in practice, instead of increasing productive efficiency, rate of return regulation often has the opposite effect. This type of intervention has the unintended consequence of encouraging utility companies to raise costs so that raising prices does not lead to higher profits and hence they comply with the rate of return regulation. The higher costs, which are the result of the productively inefficient use of resources, may allow the managers of the business to have an easier life. Monopolies are especially likely to allow costs to rise when protected by high entry barriers.
- **State ownership of monopoly.** In the past, UK Labour governments have sometimes regarded the problem of monopoly as resulting solely from private ownership and the pursuit of private profit. At its most simplistic, this view leads to the conclusion that the problem of monopoly disappears when the firms are nationalised or taken into public ownership. Once in public ownership, the monopolies are assumed to act solely in the public interest.
- **Privatising monopolies.** Opposing state ownership, past Conservative governments argued that state ownership produces particular forms of abuse that would not be experienced if the industries were privately owned. These include a general inefficiency and resistance to change, which stem from the belief by workers and management in the state-run monopolies that they will always be bailed out by government in the event of a loss. According to the Conservative view, monopoly abuse occurs in nationalised industries, not from the pursuit of private profit, but because the industries are run in the interest of a workforce that enjoys very good working conditions and is protected from market discipline. It is also alleged that interference by governments for social and political reasons made it difficult for nationalised industries to operate efficiently. The Conservatives believe that the privatisation of state-owned monopolies should improve efficiency and commercial performance, because privatisation exposes the industry to the threat of takeover and the discipline of the capital market. In some cases, privatisation has also subjected previously state-owned monopolies to competition.
- **Deregulation and the removal of barriers to entry.** Most economists believe that privatisation alone cannot eliminate the problem of monopoly abuse; it merely changes the nature of the problem to private monopoly and the commercial exploitation of a monopoly position. In recognition of this problem, the privatisation of the telecommunication, gas and electricity monopolies was accompanied by the setting up of the regulatory bodies Ofcom and Ofgem, operating in addition to the CMA. One method of exposing monopolies — including the privatised utility industries — to increased competition is to use deregulatory policies to remove artificial barriers to entry. We explain deregulation in greater detail later in the chapter.

SYNOPTIC LINK

In Chapter 5, section 5.10, we explained that in recent years, contestable market theory has had a major impact on UK monopoly policy. We described how the theory implies that, provided there is adequate potential for competition, a conventional regulatory policy is unnecessary. The government's main role should be to discover which industries and markets are potentially contestable. The growing influence of contestable market theory has been accompanied by the introduction of deregulatory policies that are meant to develop conditions in which there are reduced barriers to entry and exit or a complete absence of such barriers.

For a market to be perfectly contestable, barriers to entry must disappear and there must be no sunk costs. As we explained in Chapter 5, sunk costs are costs incurred, when entering a market, that are irrecoverable should the firm decide to leave the market.

Merger policy

Whereas a government's monopoly policy deals with established monopoly, or markets already dominated by large firms, merger policy is concerned with takeovers and mergers that might create a new monopoly. Strictly, a merger involves the voluntary coming together of two or more firms, whereas a takeover is usually involuntary, at least for the victim being acquired through a hostile takeover. However, the term 'merger policy' is used to cover all types of acquisition of firms, friendly or hostile, willing or unwilling.

Until quite recently, the government itself decided whether to refer merging companies to the competition authorities for investigation. This laid the government open to the criticism that, when deciding against a merger reference, it was bending to the lobbying power of big business and engaging in political opportunism. However, the CMA now makes virtually all merger references. The CMA keeps itself informed of all merger situations that might be eligible for investigation on public interest grounds. Currently, a takeover or merger is eligible for investigation if it is expected to lead to a substantial lessening of competition (SLC).

Restrictive trading practices policy

Restrictive trading practices undertaken by firms in imperfect product markets can be divided into two broad kinds: those undertaken independently by a single firm, and collective restrictive practices that involve either a written or an implied agreement among two or more firms.

Independently undertaken restrictive practices include:

- decisions to charge discriminatory prices (see section 5.8)
- the refusal to supply a particular resale outlet
- full-line forcing, whereby a supplier forces a distributor that wishes to sell one of its products to stock its full range of products.

The most common example of a collective restrictive trading practice is a cartel agreement, in which firms come together to fix or rig the price of a good.

Both independently undertaken and collective restrictive trading practices are now dealt with by the Competition and Markets Authority. The CMA usually asks the firm or firms involved to drop the practice voluntarily on the grounds that it is anti-competitive. A cartel agreement is usually banned, unless the firms involved persuade the CMA that the agreement is in the public interest — for example, to protect the public from injury.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.11

European Competition Commissioner fines Google €4.3 billion

In July 2018, the European Competition Commissioner Margrethe Vestager fined Google €4.3 billion for abusing its market dominance with the Android operating system, which powers most of the world's mobile phones. Vestager said:

> Our case is about three types of restrictions that Google has imposed on Android device manufacturers and network operators to ensure that traffic on Android devices goes to the Google search engine. In this way, Google has used Android as a vehicle to cement the dominance of its search engine. These practices have denied rivals the chance to innovate and compete.
>
> In particular, Google:
>
> - has required manufacturers to pre-install the Google Search app and browser app (Chrome), as a condition for licensing Google's app store (the Play Store);
> - made payments to certain large manufacturers and mobile network operators on condition that they exclusively pre-installed the Google Search app on their devices; and
> - has prevented manufacturers wishing to pre-install Google apps from selling even a single smart mobile device running on alternative versions of Android that were not approved by Google.



Internet giant Google has been accused of abusing its market dominance

Google contested the EU's decision to fine the company but lost the appeal in 2021. Google's appeal was based on the argument that Google could not possibly be a harmful monopoly because, instead of charging a price for Android, Google gives most of the operating system away. For Google's business to work, it must become as easy as possible for advertisers to reach smartphone users. That is the purpose of all the software that Google gives away, from the Android operating system, through to YouTube, Google search on phones and the Chrome browser. This might look like a cross-subsidy, but on the other hand it is the heart of Google's business model. The software that Google gives away is designed to draw phone users into Google's advertising ecosystem.

Follow-up questions

1. Find out how Google has grown since the business was founded in 1998.
2. Microsoft, the dominant company in the personal computer era, and Google, the dominant company in the internet era, have both benefited greatly from cross-subsidy and from network effects. Find out what is meant by a network effect and by cross-subsidy.
3. Do you believe that Google is a harmful monopoly? Justify your reasoning.

SECTION 8.7 SUMMARY

- Monopoly leads to the manipulation of consumers and the exploitation of producer sovereignty.
- By restricting output and raising prices, monopoly results in a net welfare loss as well as a transfer of consumer surplus into producer surplus and monopoly profit.
- Competition policy is the part of government economic policy which tries to make the imperfectly competitive and monopolistic markets of the real world more competitive.
- The aims of competition policy include preventing the exploitation of monopoly power, reducing costs of production, improving efficiency, getting rid of excessive profit so that prices reflect costs of production, and removing entry and exit barriers which separate markets.
- The Competition and Markets Authority (CMA) is the government agency responsible for advising on and implementing UK competition policy.
- The three parts of UK competition policy are monopoly policy, mergers policy and restrictive trading practices policy.
- The CMA adopts a cost–benefit approach to competition policy.

8.8 Public ownership, privatisation, regulation and deregulation of markets

Public ownership

KEY TERM
public ownership ownership of industries, firms and other assets such as social housing by central government or local government. The state's acquisition of such assets is called nationalisation.

Publicly owned or state-owned industries are also known as *nationalised industries*. Although there were nationalised industries in the UK in the nineteenth and early twentieth centuries, the main period of nationalisation occurred in the years after the Second World War. From 1945 until 1951 and then intermittently from 1964 until 1979, whenever Labour governments had control of the UK economy, industries such as coal mining, the railways and the steel industry were taken into **public ownership**. Labour governments justified nationalisation on the grounds that effective state planning required public ownership of the 'commanding heights' of the economy. Public ownership was essential for efficient operation of key industries which were regarded as too important to be left to the vagaries (unpredictability) of private ownership and market forces.

As well as being an instrument of socialist planning and control of the economy, public ownership was used (as we explained in the previous section) as a method of regulating the problem of monopoly — in particular, the problem of natural monopoly in the utility industries.

Since 1979 very few new industries have been taken into public ownership in the UK. However, there have been two examples of *temporary* nationalisation.

The first of these was the complete or partial nationalisation of the Northern Rock, Lloyds, RBS and HBOS banks. This was a response to the financial crisis which hit the UK economy (and also the wider world economy) in 2007 and 2008. Because of the extreme adverse effects that bank collapses would have on other industries and consumers, banks were regarded as 'too important to fail'. Those that were in danger of failing were either nationalised or merged. The decision to take banks into public ownership was made for pragmatic rather than ideological reasons. The intention was always, particularly under the Conservative-led Coalition government, to sell the nationalised banks back to the private sector as soon as they became financially viable.

The other industry taken into temporary public ownership was the east coast railway service. The nationalisation of this service and its eventual return to the private sector, and subsequent renationalisation, are discussed in Application of Economics in the Real World 8.12.

TEST YOURSELF 8.12
Do you agree that public ownership leads to the creation of state monopolies? Justify your argument.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.12

East Coast Rail returns to private hands, before being renationalised again

In March 2015, the *Guardian* newspaper reported that train services on the mainline between London and Edinburgh were once again being operated by a private firm after more than five years under state control, reigniting the row over ownership of the railway.

East Coast trains were rebranded as Virgin Trains East Coast. Under the management of a joint venture between two private-sector companies, Stagecoach and Virgin, a private-sector organisation took over the running of the railway on an eight-year franchise.

The previous private operators of the east coast railway line had failed to meet their financial commitments. A small government-owned company, Directly Operated Railways, stepped in to run trains on the mainline in late 2009 after the previous franchisee, National Express, walked away when revenues fell during the financial crisis.

In its five years as East Coast, the state-run firm gave several millions of pounds in profits to the Treasury. In the last two years in which it operated, East Coast was one of two state-owned firms to make a net contribution to government funds, paying in more to the government than it received in subsidy or indirect grants.

The success of the nationalised East Coast trains led to calls to keep the line in state hands. The general secretary of the Trades Union Congress (TUC), Frances O'Grady, said that sale to the private sector would be a 'costly mistake'. She said: 'By taking East Coast out of public ownership the government is passing the profits to Stagecoach and Virgin shareholders, instead of using the cash to reduce rail fares and improve services for passengers.'

However, in June 2018, the east coast rail service was temporarily renationalised because the private owners Virgin and Stagecoach could no longer meet the promised payments in the £3.3 billion contract. With a reminder of its name before nationalisation in 1948, the new operator has been rebranded as the London and North Eastern Railway.

Follow-up questions

1 When was Virgin Trains East Coast set up?
2 What were the reasons for the Conservative government granting the company a franchise to operate trains?
3 What happened to the company in June 2018?
4 Do you agree that other privately owned operators of UK railway companies should also be renationalised? Justify your answer.

The arguments for and against the privatisation of state-owned enterprises

KEY TERM
privatisation the transfer of assets from the public sector to the private sector.

Privatisation involves the transfer of publicly owned assets to the private sector. In the UK, this has usually involved the sale to private ownership of nationalised industries and businesses that were previously owned by the state and accountable to central government. Between 1981 and 2015 more than 16 major industries were privatised and state ownership of industries was transformed from the ownership of major parts of the UK economy to something much less significant. Although the main privatisations involved the sale of nationalised firms and industries, other state-owned assets such as land and socially owned housing have also been privatised.

The case for privatisation

Arguments used to justify privatisation include:

- **Revenue raising.** Privatisation, or the sale of state-owned assets, provides the government with a short-term source of revenue, which at the height of privatisation was at least £3–4 billion a year. But obviously an asset cannot be sold off twice.
- **Reducing public spending and the government's borrowing requirement.** Since 1979, Conservative governments have aimed to reduce public spending and government borrowing. When the state successfully sold loss-making industries such as the Rover Group, public spending on subsidies fell. Government borrowing also falls if private ownership returns the industries to profitability, since corporation tax revenue is boosted and the state earns dividend income from any shares that it retains in the privatised company, providing that the privatised firms remain British.
- **The promotion of competition.** Privatisation has been justified on the grounds that it promotes competition through the break-up of monopoly. At the time of their privatisation, industries such as gas and electricity were natural monopolies. But the growth of technology-driven competition, together with the removal of barriers to entry by regulating agencies such as Ofgem, has significantly increased competition.
- **The promotion of efficiency.** For free-market economists, this is perhaps the most important justification for privatisation. Supporters of privatisation believe that public ownership gives rise to special forms of inefficiency which disappear once an industry moves into the private sector. The culture of public ownership makes nationalised industries resistant to change. Through exposure to the threat of takeover and the discipline of the capital market, the privatisation of a state-owned monopoly should improve the business's efficiency and commercial performance.
- **Popular capitalism.** The promotion of an enterprise culture was an important reason for privatisation in the UK. Privatisation extended share ownership to employees and other individuals who had not previously owned shares, and thus added to the incentive for the electorate to support the private enterprise economy. Privatisation has generally proved popular with voters, so governments, both Conservative and then Labour, saw no point in abandoning a winning programme.

The case against privatisation

Privatisation has the following possible disadvantages:

- **Monopoly abuse.** Opponents of privatisation have argued that, far from promoting competition and efficiency, privatisation increases monopoly abuse by transferring socially owned and accountable public monopolies into weakly regulated and less accountable private monopolies.
- **Short-termism wins over long-termism.** Many of the investments that need to be undertaken by the previously nationalised industries can only be profitable in the long term. There is a danger that under private ownership, such investments will not be made because company boards concentrate on the short-termism of delivering dividends to keep shareholders and financial institutions happy. Under-investment in maintaining the rail track and in technically advanced trains by the privatised railway companies is said to provide an example. However, there is a counter-argument that under public ownership, the government starved the nationalised industries of investment funds in order to keep government borrowing down.
- **Selling valuable assets simply to fund government spending.** Opponents of privatisation argue that, if a private-sector business were to sell its capital assets simply in order to raise revenue to pay for current expenditure, it would rightly incur the wrath of its shareholders. The same should be true of the government and the sale of state-owned assets. Taxpayers should not sanction the sale of capital assets owned on their behalf by the UK government to raise revenue to finance current spending on items such as wages and salaries. In reply, supporters of the privatisation programme argue that, far from selling valuable assets, privatisation merely returns the family's assets to the family: that is, from the custody of the state to direct ownership by private individuals.
- **Selling valuable assets too cheaply.** Opponents of privatisation claim that state-owned assets have been sold too cheaply. This is because the offer price of shares in newly privatised industries was normally pitched at a level which guaranteed a risk-free capital gain or one-way bet at the taxpayer's expense. Arguably, this encourages the very opposite of an enterprise economy.

Regulation and deregulation of markets

Regulation

Economic **regulation** involves the imposition of rules, controls and constraints which restrict freedom of economic action in the marketplace. There are two types of regulation: external regulation and self-regulation. Both types of regulation can be imposed by law. When this is the case, a failure to comply with a regulation means breaking the law. Court cases may follow which could lead to the imposition of fines or even imprisonment.

KEY TERM
regulation the imposition of rules and other constraints which restrict freedom of economic action.

- External regulation, as the name suggests, involves an external agency laying down and enforcing rules and restraints. The external agency may be a government department such as the Department for Business, Energy and Industrial Strategies, or a special regulatory body or agency set up by government, such the Competition and Markets Authority.

- By contrast, self-regulation or voluntary regulation involves a group of individuals or firms regulating themselves — for example, through a professional association such as the Law Society or the British Medical Association, which regulate the legal and medical professions, respectively.

Regulation and market failure

Competition can sometimes bring about a situation in which social costs and benefits are not the same as the private costs and benefits incurred and received by the people actually undertaking the market activity. As explained earlier in the chapter, the over-production of externalities such as environmental pollution, and the under-consumption of education, healthcare and other merit goods provide familiar examples of divergence between private and social costs and benefits. Governments use regulation to try to correct such market failures and to achieve a socially optimal level of production and consumption. Monopoly is also a form of market failure, and regulation is used to limit and deter monopoly exploitation of consumers.

Other examples of government regulation aimed largely at reducing the social costs of market activity include health and safety at work, anti-discrimination and safeguards of workers' rights, and consumer protection legislation. Much of this regulation is concerned with the adequate provision of information for customers and workers, and the setting of quality standards for the production of goods. Such regulation may affect advertising standards, consumers' rights to replacement or repair when purchasing faulty goods, and workers' rights in the event of discrimination or unfair dismissal.

Advantages of government regulation: a summary

In a narrow sense, regulation is necessary to protect:

- consumers from harmful products and to maintain quality standards
- workers from labour market exploitation
- the environment
- children and old people from exploitation and abuse
- people from self-harm.

More broadly, regulation is necessary to enable markets to perform well and to limit the damage caused by market failures.

STUDY TIP

While there is a strong case for removing many regulations, such as those that unnecessarily raise costs of production and consumer prices, remember that many regulations can be justified on the grounds that they protect people from, for example, the abuse of monopoly power and harmful externalities. This is an important point that you could make in evaluation in a question about the advantages and disadvantages of reducing regulation.

Deregulation

KEY TERM

deregulation the removal of previously imposed regulations.

Deregulation involves the removal of any previously imposed regulations that have adversely restricted competition and freedom of market activity. During the last 35 years, significant deregulation has taken place in the UK and the USA. Systems of regulation built up during the early post-1945 period have on occasion been completely abandoned, while in other cases they have

TEST YOURSELF 8.13
With examples of each, distinguish between deregulation and privatisation.

been weakened or modified. The UK government has removed the protected legal monopoly status enjoyed, for example, by bus companies, airlines and commercial television and radio companies. Access to BT's distribution network of land lines has been given to competitors in the telecommunications industry, and private power companies have been allowed to rent the services of the national electricity and gas distribution grids.

There are two main justifications of deregulation:

- the promotion of competition and market contestability through the removal of artificial barriers to market entry
- the removal of 'red tape' (seemingly unnecessary official rules) and bureaucracy which imposes unnecessary costs on economic agents, particularly businesses.

Disadvantages of government regulation: a summary

In a narrow sense, deregulation is necessary to get rid of:

- unnecessary bureaucracy and 'red tape'
- compliance costs
- interference in individual economic decision making.

More broadly, deregulation enables markets to function more efficiently and to incur lower costs. The justification of deregulation is based on the assumption that government intervention in a market economy must be kept to a minimum and that individuals know better than governments what is in their self-interest.

Deregulation and the theory of contestable markets

Much of the justification for the policies of deregulation and economic liberalisation that have been pursued in recent years has been provided by the theory of contestable markets, which was explained earlier in the chapter and in Chapter 5. Contestable market theory argues that the most effective way to promote competitive behaviour within markets is not to impose ever-more regulation upon firms and industries, but to carry out the opposite process of deregulation.

According to this view, the main function of deregulation is to remove barriers to entry, thereby creating incentives both for new firms to enter and contest the market and for established firms to behave in a more competitive way. Under the influence of the theory of contestable markets, governments have sought to remove or loosen all regulations whose main effect has been to reduce competition and to create unnecessary barriers to market entry.

Regulatory capture

Another theory that has had some influence on the trend towards deregulation is the theory of **regulatory capture**. This theory argues that regulatory agencies created by government can be 'captured' by the industries or firms they are intended to oversee and regulate. Following capture, the regulatory agencies begin to operate in the industry's interest rather than on behalf of the consumers whom they are supposed to protect. A classic case of regulatory capture occurred when the director general of Oflot, the agency which regulated the national lottery, was caught accepting free air tickets and other 'sweeteners' from one of the lottery companies he was supposed to regulate.

KEY TERM
regulatory capture occurs when regulatory agencies act in the interest of regulated firms rather on behalf of the consumers they are supposed to protect.

Regulatory capture can also occur without the regulator behaving inappropriately. For example, the inevitable close contacts between the regulator and the regulated may mean that the regulator becomes predisposed to accepting the views of the organisations being regulated rather than those of the consumer. Also, the regulator often relies on the regulated organisations to provide much of the information which it uses to make its judgements; the information supplied may be biased.

Even if regulatory capture does not take place, the supporters of deregulation argue that much regulatory activity is unnecessary and ultimately burdensome for industry and consumers. Once established, the regulators have an incentive to extend their role by introducing ever more rules and regulations, since it is in this way that they justify their pay and their jobs. Regulation acts both as an informal 'tax' upon the regulated, raising production costs and consumer prices, and also as an extra barrier to market entry, restricting competition.

TEST YOURSELF 8.14
Describe one other example of alleged regulatory capture besides the example given here.

EXTENSION MATERIAL

The regulation of the privatised utility industries

As explained, deregulatory policies have been implemented alongside privatisation in liberalising the UK economy. In the 1980s and 1990s, UK governments realised that, once industries such as telecommunications, gas, water and electricity were privatised, there was a danger they might abuse their monopoly position and exploit the consumer. For this reason, special regulatory agencies such as Ofgem, which now regulates the gas and electricity industries, were set up at the time of privatisation to act as watchdogs over the performance of the utilities in the private sector. Initially, industry-specific regulatory bodies were created, but some of these agencies have been merged and now cover more than one industry.

The establishment of regulatory agencies such as the Office of Rail and Road (ORR) at a time when governments have actively been pursuing a policy of deregulation and economic liberalisation created a rather strange paradox and a source of possible conflict. On the one hand, by setting markets free, deregulation reduces the role of the state; on the other hand, new watchdog bodies, such as Ofgem, have extended the regulatory role of government and its agencies.

However, successive governments have argued that there need be no conflict between regulation and deregulation. This is because the regulatory bodies are themselves actively involved in deregulating the industries they oversee — for example, by enforcing the removal of barriers that prevent the entry of new firms. Recent technological progress has made it increasingly possible for new firms to enter the utility industries, particularly the telecommunications industry. By introducing competition into markets that were previously dominated by established companies such as BT and British Gas, new market entrants have eroded the natural monopoly position previously enjoyed by the privatised utilities.

Supporters of the liberalisation programme hope that the watchdog agencies will prove so successful that eventually these regulatory bodies can wither away, when the markets they oversee have become sufficiently competitive. However, at best this is likely to be a long process. Although new firms are beginning to compete in the markets previously completely dominated by state-owned utilities, established companies like British Gas are still dominant. Their continuing market power means that the regulatory bodies set up at the time of privatisation must continue for the foreseeable future as a surrogate for competition. Some commentators argue that, far from withering away, the new regulatory agencies may gradually extend their powers and functions. Free-market critics of economic regulation believe that the UK regulatory system provides a classic example of a growing bureaucracy.

SECTION 8.8 SUMMARY

- Publicly owned or state-owned industries are also known as nationalised industries.
- Labour governments justified nationalisation on the grounds that effective state planning required public ownership of the 'commanding heights' of the economy.
- Public ownership was regarded as essential for efficient operation of key industries which were regarded as too important to be left to the vagaries (unpredictability) of private ownership and market forces.
- Public ownership was also used as a method of regulating the problem of monopoly — in particular, the problem of natural monopoly in the utility industries.
- Since 2009, there has been temporary nationalisation of some banks and the east coast railway service.
- Regulation is the imposition of rules and other constraints which restrict freedom of economic action.
- Governments use regulation to correct market failures so as to achieve a socially optimal level of production and consumption.
- Regulation is used to limit and deter monopoly exploitation of consumers.
- Regulatory agencies such as Ofwat have been created to regulate privatised utility industries and the railways.
- The theory of regulatory capture argues that regulatory agencies created by government can be 'captured' by the industries or firms they are intended to oversee and regulate.
- Following capture, the regulatory agencies begin to operate in the industry's interest rather than on behalf of the consumers whom they are supposed to protect.
- Deregulation is the removal of previously imposed rules deemed to harm economic performance.

8.9 Government intervention in markets

The reasons against and for government intervention in markets

To understand why governments intervene in markets in economies such as the UK, it is useful to divide economists (and politicians) into two different groups: those who believe that unregulated markets generally work well, and those who argue that markets are prone to market failure. The former group are non-interventionists who want to leave as much as possible to market forces, while the latter group believe that government intervention can make markets work better.

- **Pro-free market economists** see a market economy as a calm and orderly place in which the market mechanism, working through incentives transmitted by price signals in competitive markets, achieves a better

or more optimal outcome than can be attained through government intervention. In essence, risk-taking business owners, who will gain or lose through the correctness of their decisions in the marketplace, know better what to produce than civil servants and planners cocooned by risk-free salaries and secured pensions. And provided that markets are sufficiently competitive, what is produced is ultimately decided by the wishes of consumers, who know better than governments what is good for them. According to this philosophy, the correct economic function of government is to act as 'nightwatchman' by maintaining law and order, providing public goods and possibly merit goods when the market fails, and generally ensuring a suitable environment in which 'wealth-creating' firms can function in competitive markets, subject to minimum interference and regulation.

- **Interventionist economists**, by contrast, believe that all too often markets are uncompetitive, characterised by monopoly power and producer sovereignty, and prone to other forms of market failure. Additionally, uncertainty about the future and lack of correct market information are destabilising forces. By intervening in the economy, especially to correct market failures, the government 'knows better' than unregulated market forces. It can anticipate and counter the destabilising forces existent in markets, achieving a better outcome than is likely in an economy subject to market forces alone.

Correcting market failures

There are various methods open to a government for correcting, or at least reducing, market failures. At one extreme, the government can abolish the market, using instead the command or planning mechanism, financed from general taxation, for providing goods and services. At the other extreme, the government can try to influence market behaviour by providing information, and by exhorting and 'nudging' firms and consumers to behave in certain ways (e.g. not to use plastic bags). Between these extremes, governments can impose regulations to limit people's freedom of action in the marketplace, and use taxes and subsidies, price ceilings and price floors to alter prices in the market in order to change incentives and economic behaviour.

Government provision of public goods and merit goods

The free-rider problem which we explained on page 235 means that markets may fail to provide pure public goods such as national defence and police services. When free-riding occurs, the incentive function of prices breaks down. If goods are provided by a market, people can free-ride rather than pay a price, so the firms that are trying to sell the goods cannot make a profit. Given that there is a need for public goods, as we have seen, governments often step into the gap and provide the goods, financing the provision out of general taxation.

Just as governments discourage the production and consumption of negative externalities and demerit goods, in much the same way they try to encourage the production and consumption of positive externalities and merit goods. The government may choose to regulate or to try to change the prices of

merit goods and other goods and activities which yield external benefits. In the latter case, subsidies are often used to encourage production and consumption.

Regulation, on the other hand, can force consumers to consume merit goods. The government may require people to be vaccinated against disease and to wear seatbelts in cars and crash helmets on motorbikes.

In the UK, education is both compulsory and completely subsidised, from 2015 at least for children between the ages of 5 and 18. Low-income families would be in an impossible situation if required to pay for education as well as to send their children to school. Subsidies can, of course, be paid to private providers of education and healthcare: namely, to private schools and private hospitals. But in the UK, education and healthcare are also provided by the state, forming an important part of public spending. Nevertheless, private-sector provision is growing. One reason for growing private-sector provision of merit goods lies in the fact that state provision does not necessarily mean good-quality provision.

SYNOPTIC LINK

Reread the earlier sections of this chapter on the various forms of market failure. This will help to remind you of the specific problems which government intervention is intended to tackle.

Forcing firms and consumers to generate positive externalities

The state can impose regulations that force firms and consumers to generate positive externalities. Local authority bylaws can require households to maintain the appearance of properties, and the state may order landowners to plant trees. In this situation, it is illegal *not* to provide external benefits for others.

Government intervention, negative externalities and demerit goods

There are two main ways in which governments can intervene to try to correct the market failures caused by negative externalities and demerit goods. As with positive externalities and merit goods, the government can use regulation and quantity controls. Alternatively, or in addition, it can use taxation.

Regulation and the correction of market failures

Governments use regulations, including compulsory consumption, and subsidies to enforce or encourage production of positive externalities and the consumption of merit goods.

Regulation can directly influence the quantity of the externality that a firm or household can generate, and the level of consumption of a demerit good such as tobacco. In its most extreme form, regulation can be used to ban completely, or criminalise, the generation of negative externalities such as pollution or the sale and consumption of a demerit good such as

heroin. However, it may be impossible to produce a good or service such as electricity in a coal-burning power station without generating at least some of a negative externality. In this situation, banning the externality has the perverse effect of preventing production of a good (e.g. electricity) as well as the bad (pollution). Because of this, quantity controls that fall short of a complete ban may be more appropriate. These include maximum emission limits and restrictions on the time of day or year during which the negative externality can legally be produced. In the case of 'milder' demerit goods, smoking can be banned in public places, while shops would break the law by selling alcohol to younger teenagers.

Taxation and the correction of market failures

Until recently, governments have been more likely to use regulation rather than taxation to reduce negative externalities such as pollution and congestion. Indeed, in the past, it was difficult to find examples of pollution taxes outside the pages of economics textbooks, possibly because politicians feared that pollution taxes would be too unpopular. But in recent years, governments have become more prepared to use congestion and pollution taxes. This reflects growing concern, among governments and the public alike, of environmental issues such as global warming and the problems posed by fossil fuel emissions and other pollutants. It may also reflect both the growing influence of green or environmental pressure groups, such as Friends of the Earth, and a growing preference to tackle environmental problems with market solutions rather than through regulation.

KEY TERM

tax a compulsory levy imposed by the government to pay for its activities. Taxes can also be used to achieve other objectives, such as reduced consumption of demerit goods.

Completely banning negative externalities and demerit goods is a form of market replacement rather than market adjustment. By contrast, because **taxes** placed on goods affect incentives which consumers and firms face, they provide a market-orientated solution to the problems posed by negative externalities and demerit goods. Taxation compensates for the fact that there is a missing market in the externality. In the case of pollution, the government calculates the money value of the negative externality, and imposes this on the firm as a pollution tax. This is known as the 'polluter must pay' principle.

The pollution tax creates an incentive, which was previously lacking, for less of the bad to be dumped on others. By so doing, the tax internalises the externality. The polluting firm must now cover all the costs of production, including the cost of negative externalities, and include these in the price charged to customers. By setting the tax so that the price the consumer pays equals the social cost of producing another unit of the good generating the negative externality, resource allocation in the economy is improved. However, a pollution tax, like any tax, itself introduces new inefficiencies and distortions into the market, associated with the costs of collecting the tax and with creating incentives to evade the tax illegally: for example, by dumping pollution at night to escape detection. This is an example of government failure, which is explained in the final section of this chapter.

SYNOPTIC LINK

To analyse how the imposition of a tax on a demerit good such as tobacco affects consumption of the good, we must make use of the concept of price elasticity of demand. Figure 8.12 shows that imposing an expenditure tax on a good in fairly elastic demand is effective in reducing demand for the product. However, because of their addictive properties, the demand for demerit goods such as alcohol, tobacco and hard drugs can be inelastic. In that case, taxing demerit goods can raise lots of revenue for the government, but not do much to reduce consumption. And if the tax is set at a very high rate, it may lead to smuggling and to black market activity.

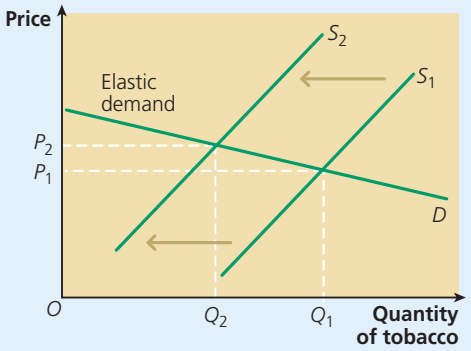


Figure 8.12 The effect of imposing a tax on tobacco when demand is elastic

STUDY TIP

Make sure you can analyse the effects of taxes imposed on goods, and of subsidies, by applying the concept of price elasticity of demand. This will help you to understand the effectiveness of taxes and subsidies in moving the equilibrium quantity consumed and produced towards the optimum level.

Pollution permits

Until quite recently, the main choice of policy for dealing with the problem of pollution was between regulation and taxation. As we have explained, the former is an interventionist solution whereas taxation, based on the principle that the polluter must pay, has been seen as a more market-orientated solution, but nevertheless one which required the government to levy and collect the pollution tax. In the 1990s, another market-orientated solution started in the USA, based on a trading market in permits or licences to pollute. More locally to the UK, the EU Emissions Trading Scheme is now the centrepiece of European efforts to cut emissions.

A 'permits to pollute' scheme (for electricity) still involves regulation: for example, the imposition of maximum limits on the amount of pollution that coal-burning power stations are allowed to emit, followed by a steady reduction in these ceilings in each subsequent year (say, by 5%). But once this regulatory framework has been established, a market in traded pollution permits takes over, creating market-orientated incentives for the power station companies to reduce pollution because they can make money out of it.

TEST YOURSELF 8.15
Research how the sale of pollution permits has been used in the USA to try to reduce environmental pollution.

A tradable market in permits to pollute works in the following way. A price is set in the market for permits by the interaction of demand for, and supply of, the permits. Energy companies able to reduce pollution by more than the law requires sell their spare permits to other power stations that, for technical or other reasons, decide not to, or cannot, reduce pollution below the maximum limit. The latter still comply with the law, even when exceeding the maximum emission limit, because they buy the spare permits sold by the first group of power stations. But, in the long run, even power stations that find it difficult to comply with the law have an incentive to reduce pollution, so as to avoid the extra cost of production created by the need to buy pollution permits.

Price ceilings or maximum price laws

KEY TERM
price ceiling a price *above* which it is illegal to trade. Price ceilings, or maximum legal prices, can distort markets by creating excess demand.

Perhaps the simplest way in which a government can impose a price control is through the use of a **price ceiling** or a price floor. Suppose, for example, that in a particular market — say, the market for bread — the government imposes a price ceiling or maximum legal price, shown as P_1 in Figure 8.13. Because the price ceiling has been imposed below the free-market equilibrium price of P^*, it creates excess demand, shown by the distance between Q_1 and Q_2. In a free market, market forces would raise the price and eliminate the excess demand. But, because the price ceiling prevents this happening, there is no mechanism in the market for getting rid of excess demand. Rather than rationing by price, households are rationed by quantity. Queues and waiting lists occur, and possibly bribery and corruption, through which favoured customers buy the good but others do not. It is also worth noting that price ceilings interfere with the incentive function of prices, in the sense that the ceilings prevent prices from rising to attract new firms into the market.

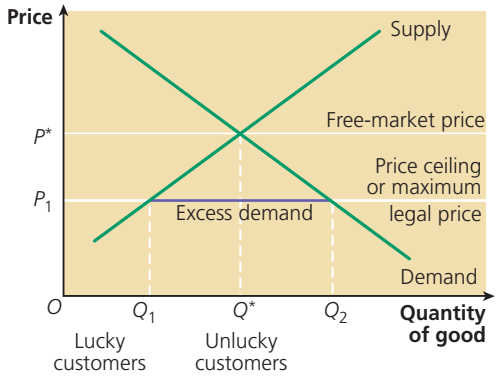


Figure 8.13 The effect of a maximum price control or price ceiling

The emergence of an informal or shadow market (sometimes called a black market) is also likely. Secondary markets emerge when primary markets (or free markets) are prevented from working properly. A secondary market is a meeting place for lucky and unlucky customers. In the secondary market, some lucky customers, who bought the good at price P_1, resell at a higher price to unlucky customers unable to purchase the good in the primary market.

Price floors or minimum legal prices

KEY TERM
price floor a price *below* which it is illegal to trade. Price floors, or minimum legal prices, can distort markets by creating excess supply.

Sometimes governments impose minimum price laws or **price floors**. For a minimum price law to affect a market, the price floor must be set above the free-market price. Figure 8.14 illustrates the possible effect of the national minimum wage imposed in UK labour markets. A national minimum wage rate set at W_1 (which is *above* the free-market wage rate of W^*) creates an excess supply of labour, thereby causing unemployment equal to the distance between L_1 and L_2. It may also cause rogue employers to break the law: for example, paying 'poverty wages' to vulnerable workers such as illegal immigrants. Note also that whereas a price ceiling imposed above the free-market price in Figure 8.13 would have no effect on the price at which bread is traded in the market, a national minimum wage set below the free-market wage rate in Figure 8.14 would have no effect on unemployment. This is the situation in many UK labour markets.

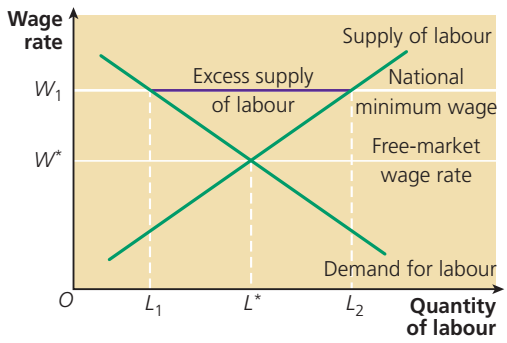


Figure 8.14 The possible effect of the UK national minimum wage

As with price ceilings, price floors interfere with the incentive function of prices. This is because falling prices cause inefficient or high-cost firms to leave the market. A price floor prevents this from happening.

TEST YOURSELF 8.16

With the help of appropriate diagrams, relate the effects of maximum and minimum legal prices to the concepts of excess demand and excess supply.

STUDY TIP

Make sure you can describe and explain at least three ways in which government intervention can affect the price of a good or service. This will help you to produce an extended response to a question on methods of intervention to tackle market failure.

QUANTITATIVE SKILLS 8.5

Worked example: calculating the effect of a price ceiling

Figure 8.15 shows how the imposition of a price ceiling distorts a market.

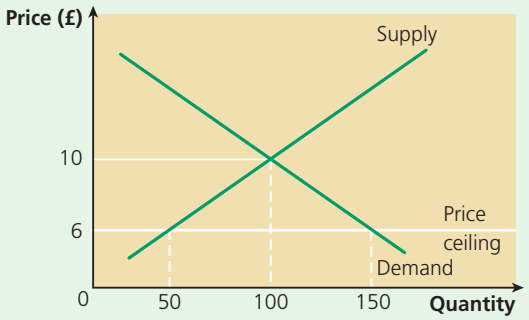


Figure 8.15 Price ceiling

Calculate the total amount spent on the good, first at market equilibrium price, and then when the price ceiling is in place.

At any price which might be charged for the good, the money value of the good bought and sold is calculated by multiplying price by quantity ($P \times Q$). At the good's equilibrium or market-clearing price of £10, 100 units of the good are bought and sold, so the amount spent on the good is £1,000. However, when a price ceiling of £6 is imposed, only 50 units of the good are bought and sold. The money spent on the good falls to £300, even though consumers would like to buy 150 units of the good at this price.

SECTION 8.9 SUMMARY

- Pro-free-market and interventionist economists hold different views on when governments should intervene in markets.
- Reasons for intervention include the correction of market failures, which includes government provision of public goods and merit goods and forcing or incentivising firms and consumers to generate positive externalities.
- Methods of intervention include regulation and taxation, the imposition of price ceilings and price floors, and the introduction of pollution permits.

8.10 Government failure

KEY TERM

government failure occurs when government intervention reduces economic welfare, leading to an allocation of resources that is worse than the free-market outcome.

When explaining market failure, we assumed that it can be reduced or completely eliminated, once identified, through appropriate government intervention: for example, by imposing taxes, controls or regulation. But there is another possibility. When the government intervenes in the economy to reduce or correct market failure, its intervention may lead to the appearance of other forms of resource misallocation. When this happens, **government failure** occurs.

We saw earlier that governments often provide free state education because they wish to increase consumption of what they perceive to be a merit good.

If you refer back to Figure 8.10, you will see that the socially optimal level of provision of education, Q_2, is higher than Q_1, the amount of education which markets would probably provide. However, government intervention does not in itself ensure that Q_2 is eventually provided. Since it makes decisions on the basis of imperfect information, the government may not know what is the socially optimal level of education. Some people argue that free state education leads to over-provision because people want to consume the quantity available when zero price is charged for education. If this is the case, governments may create, rather than remove, market distortions. Some people also believe that the quality of state education is poorer than that of private market-based education. If this is the case, it is a further example of government failure.

Reasons for government failure

The pursuit of conflicting policy objectives

A good example is provided by government investment decisions. When governments wish to expand the economy, they may give the go-ahead to long-term investment projects such as the building of new roads, which they then cancel because of the perceived need to contract the economy. Very often government decisions can be criticised for favouring short-termism at the expense of long-termism.

Administrative costs

There are also administrative costs to consider. Government intervention in the economy to correct market failure may create unnecessary layers of bureaucracy, which create costs which taxpayers have to pay. If so, this is a government failure. Examples are the high costs the UK government incurred

when deciding whether or not to permit the construction of a third runway at Heathrow airport, or to expand Gatwick or Stansted, or to go ahead with the High Speed 2 (HS2) railway line. Many free-market economists believe this is unnecessary government expenditure, but more interventionist economists believe the expenditure is fully justified because it leads to better eventual government decisions.

The 'law of unintended consequences'

The 'law of unintended consequences', which has become very fashionable in recent years, predicts that, whenever the government intervenes in the market economy, effects will be unleashed which the policy-makers had not foreseen or intended. Sometimes the unintended effects may be advantageous to the economy, while in other instances they may be harmful but relatively innocuous. In either of these circumstances, government intervention can be justified on the grounds that the social benefits of intervention exceed the social costs and therefore contribute to a net gain in economic welfare. But if government activity — however well intentioned — triggers harmful consequences which are greater than the benefits that the government intervention is supposed to promote, then government failure results.

TEST YOURSELF 8.17
With examples of each, explain the difference between government failure and market failure.

STUDY TIP
It is important to avoid confusing *government* failure with market failure if asked about it in your exam. Government failure can be the result of government intervention to correct cases of market failure.

Examples of possible government failure

Government price controls create black markets

In section 8.9, we explained how a price ceiling or maximum price law can create excess demand in a market, which is then relieved through trading in an informal or shadow market: in other words, a black market. Price ceilings are normally put in place to protect consumers from high prices. However, the rising price of a product may simply reflect market forces and the changing nature of supply or demand in the market. A higher price might be needed to create incentives for consumers to economise and for firms to divert more scarce resources into producing the good. The price ceiling may prevent this happening. The controlled price can send out the wrong signals and create the wrong incentives, thus contributing to resource misallocation. And since it may be a criminal activity to break the price law, black markets are sometimes characterised by corruption and the threat of the use of illegal force.

However, economists of a free-market persuasion often justify black markets on the grounds that they do the job that the primary market should do: that is, equate demand with supply. A price ceiling prevents the primary market from working properly. Arguably, the touts and dealers who act as middlemen in the black market or underground economy contribute to better resource allocation, although their contribution would not be needed if there were no price controls. A black market or secondary market only comes into existence because price controls distort the primary market.

APPLICATION OF ECONOMICS IN THE REAL WORLD 8.13

The landfill tax and government failure

Government policies that aim to reduce the discharge of negative externalities can also lead to government failure. Almost every economic activity produces waste: for example, household rubbish and the waste created by building and construction. A large proportion of UK waste is either incinerated (which discharges pollutants into the atmosphere) or collected by local government and placed in landfill sites. Landfill also causes pollution by releasing gases such as carbon dioxide and methane, and a further problem arises as all the available landfill sites fill up.

In 1996, the UK government imposed a landfill tax which it hoped would create jobs and reduce waste. But to evade the tax, rogue building contractors and some households began to fly-tip and to dump rubbish in public places and on other people's land. This was an unintended and adverse consequence of a tax that was intended to improve the environment.



Government policies to reduce landfill have had the unintended consequence of increasing illegal fly-tipping

Many blame the controversial landfill tax for the rise in organised unauthorised dumping. The tax increased the costs of taking waste to licensed sites by up to a third. The cost of getting rid of one truckload of rubble could be as high as £400. Finding alternative dumping grounds, where off-loading a lorry costs little or nothing, allows unscrupulous people to make a fortune. But the cost to the environment is immense.

Research published in 2022 by the Countryside Alliance uncovered the enormous scale of illegal fly-tipping in England and Wales. Illegal fly-tipping can cost taxpayers over £392 million per year. At least 1.13 million incidents of unlawful rubbish dumping were recorded in England and Wales between April 2020 and March 2021, which works out at almost 130 incidents of fly-tipping every hour — more than two every minute!

Typically only one in 50 cases leads to a prosecution. In cash-strapped rural local authorities, the rate of prosecutions dropped to just three in every 1,000. If waste is dumped on private land, the owners, irrespective of having no part in the fly-tip, have a duty of care and are bound by law to clear it up in their own time and at their expense.

The chief executive of the Countryside Alliance said, 'With the government raising the landfill tax and with more cuts coming to council budgets, this problem is only going to get worse.'

Follow-up questions

1. Relate fly-tipping to the concept of negative externalities.
2. Do you agree that landfill illustrates both market failure and government failure? Justify your answer.
3. In Singapore, littering fines range from S$300 to S$1,000 for first-time offenders. The fine rises up to S$5,000 for third-time offenders with the possibility of facing time in prison. Do you think that similar fines should be imposed in the UK on people who are caught fly-tipping?
4. How might nudge theory be used to discourage people from dropping litter or fly-tipping?

QUANTITATIVE SKILLS 8.6

Worked example: calculating price elasticity of demand

Between 2018 and 2022 a government increased its tax on waste disposal. Table 8.8 shows the changing price of waste disposal at a landfill site.

Table 8.8 Price of waste disposal

Year	£ per tonne of waste
2018	10
2019	11
2020	12
2021	13
2022	14

If the quantity of landfill space demanded fell by 10% between 2018 and 2022, calculate from the data the price elasticity of demand for landfill space.

If the quantity demanded fell by 10% and the price per tonne rose by 40%, the price elasticity of demand for landfill space was:

$$-10\%/+40\%$$

which is −0.25. According to the data, since the elasticity statistic (ignoring the minus sign) was less than 1 or unity, the demand for landfill space was inelastic with respect to the change in price.

SECTION 8.10 SUMMARY

- Governments use regulations, including prohibition, and taxation to prevent or reduce production of negative externalities and to reduce consumption of demerit goods.
- Governments use regulations, including compulsory consumption, and subsidies to enforce or encourage production of positive externalities and consumption of merit goods.
- Along with public goods, merit goods such as education are often provided by governments.
- A price ceiling imposed below the free-market price distorts the market and creates excess demand.
- Governments impose price floors or minimum legal prices, such as the national minimum wage, to prevent prices falling below desired levels.
- Governments may create, rather than remove, market distortions.
- Government failure occurs when government intervention in markets fails to correct market failure and/or leads to outcomes worse than the intervention was meant to correct. Like market failure, government failure is associated with a misallocation of resources.
- Government failure can result from government decisions made on the basis of inadequate information; as a result of conflicting objectives; and from the administrative costs of government intervention. It is also associated with the unintended consequences of government intervention in markets.

Questions

1 Do you agree that government policies which aim to correct market failure are always successful? Justify your answer.

2 Evaluate different forms of government intervention to deal with the problems caused by negative externalities.

3 Evaluate the view that, if merit goods are provided free by the state, the socially optimal level of consumption is always achieved.

4 Explain how monopoly may lead to market failure.

5 Using the concept of elasticity, explain why taxing demerit goods may be relatively ineffective in reducing their consumption.

6 With the help of an appropriate diagram, explain how a maximum legal price may distort a market.

7 Using examples, explain the difference between market failures resulting from inefficiency and those resulting from inequity.

8 Do you agree that markets can be relied upon to reduce the problems of environmental pollution? Justify your answer.

9 Evaluate the advantages and disadvantages of privatisation.

Microeconomics key terms

abnormal profit (also known as **supernormal profit** and **above-normal profit**) profit over and above normal profit.

absolute poverty a condition characterised by severe deprivation of basic human needs, including food, safe drinking water, sanitation facilities, health, shelter, education and information. It depends not only on income but also on access to services.

allocative efficiency occurs when it is impossible to improve overall economic welfare by reallocating resources between markets. In the whole economy, price must equal marginal cost ($P = MC$) in every market.

allocative function of prices changing relative prices allocate scarce resources away from markets exhibiting excess supply and into markets in which there is excess demand.

allocative inefficiency occurs when $P > MC$, in which case too little of a good is produced and consumed, and when $P < MC$, in which case too much of a good is produced and consumed.

altruism concern for the welfare of others.

anchoring a cognitive bias describing the human tendency when making decisions to rely too heavily on the first piece of information offered (the so-called 'anchor'). Individuals use an initial piece of information when making subsequent judgements.

artificial barriers 'man-made' barriers to market entry (e.g. patent protection).

asymmetric information when one party to a market transaction possesses less information relevant to the exchange than the other.

availability bias occurs when individuals make judgements about the likelihood of future events according to how easy it is to recall examples of similar events.

average cost of labour total wage costs divided by the number of workers employed.

average fixed cost total cost of employing the fixed factors of production to produce a particular level of output, divided by the size of output: $AFC = TFC \div Q$.

average returns of labour total output divided by the total number of workers employed.

average revenue total revenue divided by output.

average total cost (also known as **average cost**) total cost of producing a particular level of output, divided by the size of output: $ATC = AFC + AVC$.

average variable cost total variable cost divided by the size of output.

behavioural economics a method of economic analysis that applies psychological insights into human behaviour to explain how individuals make choices and decisions.

bounded rationality when making decisions, individuals' rationality is limited by the information they have, the limitations of their minds, and the finite amount of time available in which to make decisions.

bounded self-control where individuals have limited self-control to act rationally in their own interests.

capital good (also known as a **producer good**) a good which is used in the production of other goods or services.

capital productivity output per unit of capital.

cartel a collusive agreement by firms, usually to fix prices. Sometimes there is also an agreement to restrict output and to deter the entry of new firms.

choice choosing between alternatives when making a decision on how to use scarce resources.

choice architecture a framework setting out different ways in which choices can be presented to consumers, and the impact of that presentation on consumer decision making.

cognitive bias a systematic error in thinking that affects the decisions and judgements that people make.

collective bargaining a process by which wage rates and other conditions of work are negotiated and agreed upon by a union or unions with an employer or employers.

command economy (also known as a **planned economy**) an economy in which government officials or planners allocate economic resources to firms and other productive enterprises.

Competition and Markets Authority (CMA) government agency responsible for advising on and implementing UK competition policy.

competition policy the part of the government's microeconomic policy and industrial policy which aims to make goods markets more competitive. It comprises policy towards monopoly, mergers and restrictive trading practices.

competitive demand when a good is viewed by consumers as an alternative for another good, i.e. the two goods are substitutes.

competitive markets markets in which the large number of buyers and sellers possess good market information and can easily enter or leave the market.

complementary goods when two goods are complements, they experience joint demand.

complete market failure a market fails to function at all and a 'missing market' results.

composite demand demand for a good which has more than one use, which means that an increase in demand for one use of the good reduces the supply of the good for an alternative use. It is related to the concept of competing supply.

concentration ratio measures the market share (percentage of the total market) of the biggest firms in the market. For example, a five-firm concentration ratio measures the aggregate market share of the largest five firms.

condition of demand a determinant of demand, other than the good's own price, that fixes the position of the demand curve.

condition of supply a determinant of supply, other than the good's own price, that fixes the position of the supply curve.

constant returns to scale when the scale of all the factors of production employed increases, output increases at the same rate.

consumer good a good which is consumed by individuals or households to satisfy their needs or wants.

consumer surplus a measure of the economic welfare enjoyed by consumers: surplus utility received over and above the price paid for a good.

consumption externality an externality (which may be positive or negative) generated in the course of consuming a good or service.

contestable market a market in which the potential exists for new firms to enter the market. A perfectly contestable market has no entry or exit barriers and no sunk costs, and both incumbent firms and new entrants have access to the same level of technology.

creative destruction capitalism evolving and renewing itself over time through new technologies and innovations replacing older technologies and innovations.

cross elasticity of demand measures the extent to which the demand for a good changes in response to a change in the price of another good; it is calculated by dividing the percentage change in quantity demanded by the percentage change in the price of another good.

deadweight loss the loss of economic welfare when the maximum attainable level of total welfare fails to be achieved.

decrease in demand a leftward shift of the demand curve.

decrease in supply a leftward shift of the supply curve.

decreasing returns to scale when the scale of all the factors of production employed increases, output increases at a slower rate.

default choice an option that is selected automatically unless an alternative is specified.

demand the quantity of a good or service that consumers are willing and able to buy at given prices in a given period of time.

demerit goods goods, such as tobacco, for which the social costs of consumption exceed the private costs. Value judgements are involved in deciding that a good is a demerit good.

deregulation the removal of previously imposed regulations.

derived demand demand for a good or factor of production, wanted not for its own sake, but as a consequence of the demand for something else.

diseconomies of scale as output increases, long-run average cost rises.

disequilibrium a situation in which opposing forces are out of balance.

distribution of income how income is divided between rich and poor, or between different groups in society, e.g. on a regional, age or gender basis.

distribution of wealth how wealth is divided between rich and poor, or between different groups in society, e.g. on a regional, age or gender basis.

division of labour different workers performing different tasks in the course of producing a good or service.

divorce of ownership from control the owners and those who control the firm (managers) are different groups with different objectives.

duopoly two firms only in a market.

dynamic efficiency measures improvements in productive efficiency that occur in the long run over time.

economic growth the increase in the *potential* level of real output the economy can produce over a period of time.

economic system the set of institutions within which a community decides what, how and for whom to produce.

economic welfare the economic well-being of an individual, a group within society, or an economy.

economies of scale as output increases, long-run average cost falls.

effective demand the desire for a good or service backed by an ability to pay.

elasticity the proportionate responsiveness of a second variable to an initial change in the first variable.

elasticity of demand for labour proportionate change in demand for labour following a change in the wage rate. The elasticity can be calculated by using the following formula:

percentage change in quantity of labour demanded / percentage change in the wage rate

elasticity of supply of labour proportionate change in the supply of labour following a change in the wage rate. The elasticity can be calculated by using the following formula:

percentage change in quantity of labour supplied / percentage change in the wage rate

entry barriers obstacles that make it difficult for a new firm to enter a market.

equality when everyone is treated exactly the same. A completely equal distribution of income means that everybody has the same income.

equilibrium a state of rest or balance between opposing forces.

equity when everyone is treated fairly.

excess demand when consumers wish to buy more than firms wish to sell, with the price below the equilibrium price.

excess supply when firms wish to sell more than consumers wish to buy, with the price above the equilibrium price.

exchange to give something in return for something else received. Money is a medium of exchange.

excludable good people who do not want to pay can be excluded from benefiting from the good.

exit barriers obstacles that make it difficult for an established firm to leave a market.

external diseconomy of scale an increase in long-run average costs of production resulting from the growth of the market or industry of which the firm is a part.

external economy of scale a fall in long-run average costs of production resulting from the growth of the market or industry of which the firm is a part.

externality a public good, in the case of an external benefit, or a public bad, in the case of an external cost, that is 'dumped' on third parties outside the market.

factors of production inputs into the production process (e.g. land, labour, capital and enterprise).

fairness the quality of being impartial, just, or free of favouritism. It can mean treating people equally, sharing with others, giving others respect and time, and not taking advantage of them.

firm a productive organisation which sells its output of goods and/or services commercially.

fixed cost a cost of production which, in the short run, does not change with output.

framing how something is presented (the 'frame') influences the choices people make.

free-rider problem a free-rider is someone who benefits without paying as a result of non-excludability. Customers may choose not to pay for a good, preferring instead to free-ride, with the result that the incentive to provide the good through the market disappears.

full employment when all who are able and willing to work are employed.

fundamental economic problem how best to make decisions about the allocation of scarce resources among competing uses, so as to improve and maximise human happiness and welfare.

geographical immobility of labour when workers are unwilling or unable to move from one area to another in search of work.

Gini coefficient measures the extent to which the distribution of income or wealth among individuals or households within an economy deviates from a perfectly equal distribution.

government failure occurs when government intervention reduces economic welfare, leading to an allocation of resources that is worse than the free-market outcome.

hypothesis of diminishing marginal utility for a single consumer, the marginal utility derived from a good or service diminishes for each additional unit consumed.

hit-and-run competition occurs when a new entrant can 'hit' the market, make profits and then 'run', given that there are no or low barriers to exit.

immobility of labour the inability of labour to move from one job to another, either for occupational reasons (e.g. the need for training) or for geographical reasons (e.g. the cost of moving to another part of the country).

incentive function of prices prices create incentives for people to alter their economic behaviour, e.g. a higher price creates an incentive for firms to supply more of a good or service.

income personal or household income is the flow of money a person or household receives in a particular time period.

income elasticity of demand measures the extent to which the demand for a good changes in response to a change in income; it is calculated by dividing the percentage change in quantity demanded by the percentage change in income.

increase in demand a rightward shift of the demand curve.

increase in supply a rightward shift of the supply curve.

increasing returns to scale when the scale of all the factors of production employed increases, output increases at a faster rate.

incumbent firm a business with an established position in the market which often benefits from consumer loyalty and economies of scale.

individual demand the quantity of a good or service that a particular consumer or individual is willing and able to buy at different market prices.

inferior good a good for which demand decreases as income rises and demand increases as income falls.

information failure occurs when people make wrong decisions because they do not possess or they ignore relevant information. Very often they think just of the present rather than the future consequences of their actions.

innovation improves on or makes a significant contribution to something that has already been invented, thereby turning the results of invention into a product.

internal economies and diseconomies of scale changes in long-run average costs of production resulting from changes in the size or scale of a firm or plant.

invention making something entirely new; something that did not exist before at all.

joint supply when one good is produced, another good is also produced from the same raw materials, perhaps as a by-product.

labour productivity output per worker.

law of diminishing marginal returns (also known as **law of diminishing marginal productivity**) a *short-term* law which states that, as a variable factor of production is added to a fixed factor of production, both the marginal and eventually the average returns to the variable factor will begin to fall.

long run the time period in which no factors of production are fixed and all the factors of production can be varied.

Lorenz curve a graph on which the cumulative percentage of total national income or wealth is plotted against the cumulative percentage of population (ranked in increasing size of share). The extent to which the curve dips below a straight 45-degree diagonal line indicates the degree of inequality of distribution.

mandated choice people are required, often by law, to make a decision.

marginal cost addition to total cost resulting from producing one additional unit of output.

marginal cost of labour the addition to a firm's total cost of production resulting from employing one more worker.

marginal physical product of labour the addition to a firm's total output brought about by employing one more worker.

marginal returns of labour the change in the quantity of total output resulting from the employment of one more worker, holding all the other factors of production fixed.

marginal revenue addition to total revenue resulting from the sale of one more unit of the product.

marginal revenue product of labour the money value of the addition to a firm's total output brought about by employing one more worker.

marginal tax rate the tax rate levied on the last pound of income received. The term can be applied solely to income taxes or to all the taxes that a person or business pays.

marginal utility the additional welfare, satisfaction or pleasure gained from consuming one extra unit of a good or service.

market a voluntary meeting of buyers and sellers with exchange taking place.

market conduct the pricing and marketing policies pursued by firms. This is also known as market behaviour, but is not to be confused with market performance, which refers to the end results of these policies.

market demand the quantity of a good or service that all the consumers in a market are willing and able to buy at different market prices.

market disequilibrium exists at any price other than the equilibrium price, when either planned demand < planned supply or planned demand > planned supply.

market economy an economy in which goods and services are purchased through the price mechanism in a system of markets.

market equilibrium a market is in equilibrium when planned demand equals planned supply, where the demand curve crosses the supply curve.

market failure when the market mechanism leads to a misallocation of resources in the economy, either completely failing to provide a good or service or providing the wrong quantity.

market structure the organisational and other characteristics of a market.

market supply the quantity of a good or service that all the firms in a market plan to sell at given prices in a given period of time.

merit good a good, such as healthcare, for which the social benefits of consumption exceed the private benefits. Value judgements are involved in deciding that a good is a merit good.

missing market a situation in which there is no market because the functions of prices have broken down.

mixed economy an economy that contains both a large market sector and a large non-market sector in which the planning mechanism operates.

monopolistic competition a market structure in which firms have many competitors, but each one sells a slightly different product.

monopoly one firm only in a market.

monopoly power (also known as **market power**) the ability of a monopoly to raise and maintain price above the level that would prevail under perfect competition. Market power can also be exercised, usually to a lesser degree, by firms in oligopoly and monopolistic competition.

monopsony there is only one buyer in a market.

monopsony power the market power exercised in a market by the buyer of a good or the services of a factor of production such as labour, even though the firm is not a pure monopsonist.

moral hazard the tendency of individuals and firms, once insured against some contingency, to behave so as to make that contingency more likely.

mortgage a long-term loan usually used to buy a house. The loan is secured against the value of the house.

national minimum wage a minimum wage or wage rate that must by law be paid to employees, though in many labour markets the wage rate paid by employers is above the national minimum wage.

natural barriers (also known as **innocent barriers**) barriers to market entry which are not caused by deliberate actions of firms in the industry.

need something that is necessary for human survival, such as food, clothing, warmth or shelter.

negative externality an external cost that occurs when the consumption or production of a good causes costs to a third party, where the social cost is greater than the private cost.

non-renewable resource (also known as a **finite resource**) a resource (e.g. oil) that is scarce and runs out as it is used.

normal good a good for which demand increases as income rises and demand decreases as income falls.

normal profit the minimum profit a firm must make to stay in business, which is, however, insufficient to attract new firms into the market.

normative statement a statement that includes a value judgement and cannot be refuted just by looking at the evidence.

nudges factors which encourage people to think and act in particular ways. Nudges try to shift group and individual behaviour in ways which comply with desirable social norms.

occupational immobility of labour when workers are unwilling or unable to move from one type of job to another, e.g. because different skills are needed.

opportunity cost the cost of giving up the next best alternative.

partial market failure a market does function, but it delivers the 'wrong' quantity of a good or service, which results in resource misallocation.

perfect competition a market that displays the six conditions of: a large number of buyers and sellers; perfect market information; the ability to buy or sell as much is desired at the ruling market price; the inability of an individual buyer or seller to influence the market price; a uniform or homogeneous product; and no barriers to entry or exit in the long run.

plant an establishment, such as a factory, a workshop or a retail outlet, owned and operated by a firm.

positive externality an external benefit that occurs when the consumption or production of a good causes a benefit to a third party, where the social benefit is greater than the private benefit.

positive statement a statement of fact that can be scientifically tested to see if it is correct or incorrect.

poverty the state of being extremely poor and not having enough money or income to meet basic needs.

price agreement an agreement between a firm, similar firms, suppliers or customers regarding the pricing of a good or service.

price ceiling a price *above* which it is illegal to trade. Price ceilings, or maximum legal prices, can distort markets by creating excess demand.

price discrimination charging different prices to different customers for the same product or service, with the prices based on different willingness to pay.

price elasticity of demand measures the extent to which the demand for a good changes in response to a change in the price of that good.

price elasticity of supply measures the extent to which the supply of a good changes in response to a change in the price of that good.

price floor a price *below* which it is illegal to trade. Price floors, or minimum legal prices, can distort markets by creating excess supply.

price leadership the setting of prices in a market, usually by a dominant firm, which is then followed by other firms in the same market.

price war occurs when rival firms continuously lower prices to undercut each other.

price-maker when a firm faces a downward-sloping demand curve for its product, it possesses the market power to set the price at which it sells the product.

price-taker a firm which is so small that it has to accept the ruling market price. If the firm raises its price, it loses all its sales; if it cuts its price, it gains no advantage.

private good a good, such as an orange, that is excludable and rival.

privatisation the transfer of assets from the public sector to the private sector.

producer surplus a measure of the economic welfare enjoyed by firms or producers: the difference between the price a firm succeeds in charging and the minimum price it would be prepared to accept.

product differentiation the marketing of generally similar products with minor variations or the marketing of a range of different products.

production converts inputs or factor services into outputs of goods and services.

production externality an externality (which may be positive or negative) generated in the course of producing a good or service.

production possibility frontier a curve depicting the various combinations of two products (or types of product) that can be produced when all the available resources are fully and efficiently employed.

productive efficiency for the economy as a whole occurs when it is impossible to produce more of one good without producing less of another. For a firm it occurs when the average total cost of production is minimised.

productivity output per unit of input.

productivity gap the difference in labour productivity between, for example, the UK and other developed economies.

profit the difference between total sales revenue and total cost of production.

profit maximisation occurs at the level of output at which total profit is greatest.

progressive taxation a tax is progressive when, as income rises, a greater proportion of income is paid in taxation. The term can be applied to a particular tax such as income tax or to taxation in general.

property right the exclusive authority to determine how a resource is used.

public good a good, such as a radio programme, that is non-excludable and non-rival.

public ownership ownership of industries, firms and other assets such as social housing by central government or local government. The state's acquisition of such assets is called nationalisation.

quantity-setter when a firm faces a downward-sloping demand curve for its product, it possesses the market power to set the quantity of the good it wishes to sell.

quasi-public good a good which is not fully non-rival and/or where it is possible to exclude people from consuming the product.

rational behaviour acting in pursuit of self-interest, which for a consumer means attempting to maximise the welfare, satisfaction or utility gained from the goods and services consumed.

rationing function of prices rising prices ration demand for a product.

regulation the imposition of rules and other constraints which restrict freedom of economic action.

regulatory capture occurs when regulatory agencies act in the interest of regulated firms rather on behalf of the consumers they are supposed to protect.

relative poverty occurs when income is below a specified proportion of average income, e.g. below 60% of median income.

renewable resource a resource (e.g. timber) that with careful management can be renewed as it is used.

resource allocation the process through which the available factors of production are assigned to produce different goods and services (e.g. how many of society's economic resources are devoted to supplying different products such as food, cars, healthcare and defence).

restricted choice offering people a limited number of options so that they are not overwhelmed by the complexity of the situation. If there are too many choices, people may make a poorly thought-out decision or not make any decision.

returns to scale the rate by which output changes if the scale of all the factors of production is changed.

rival good when one person consumes a private good, the quantity available to others diminishes.

rules of thumb thinking shortcuts, or informed guesses, that individuals use to make decisions in order to save time and effort.

ruling market price (also known as **equilibrium price**) the price at which planned demand equals planned supply.

satisficing achieving a satisfactory outcome rather than the best possible outcome.

scarcity results from the fact that people have unlimited wants but resources to meet these wants are limited. In essence, people would like to consume more goods and services than the economy is able to produce with its limited resources.

short run the time period in which at least one factor of production is fixed and cannot be varied.

signalling function of prices prices provide information to buyers and sellers.

social benefit the total benefit of an activity, including the external benefit as well as the private benefit. Expressed as an equation: social benefit = private benefit + external benefit.

social cost the total cost of an activity, comprising the external cost as well as the private cost. Expressed as an equation: social cost = private cost + external cost.

social norms forms or patterns of behaviour considered acceptable by a society or group within that society.

specialisation a worker only performing one task or a narrow range of tasks. Also, different firms specialising in producing different goods or services.

static efficiency efficiency (e.g. productive and allocative efficiency) at a particular point in time.

subsidy a payment made by government or other authority, usually to producers, for each unit of the subsidised good that they produce. Consumers can also be subsidised: for example, bus passes given to children to enable them to travel on buses free or at a reduced price.

substitute goods alternative goods that could be used for the same purpose.

sunk costs costs that have already been incurred and cannot be recovered.

supply the quantity of a good or service that producers are willing and able to sell at given prices in a given period of time.

tax a compulsory levy imposed by the government to pay for its activities. Taxes can also be used to achieve other objectives, such as reduced consumption of demerit goods.

technical progress new and better ways of making goods and new techniques for producing more output from scarce resources.

technological change a term used to describe the overall effect of invention, innovation and the diffusion or spread of technology in the economy.

total cost all the cost incurred when producing a particular size of output.

total returns the whole output produced by all the factors of production, including labour, employed by a firm.

total revenue all the money received by a firm from selling its total output.

trade the buying and selling of goods and/or services.

trade union a group of workers who join together to maintain and improve their conditions of employment, including their pay.

unemployment when not all of those who are able and willing to work are employed.

utility the satisfaction or economic welfare an individual gains from consuming a good or service.

variable cost a cost of production which changes with the amount that is produced, even in the short run.

wage discrimination paying different workers different wage rates for doing the same job.

want something that is desirable, such as a mobile phone, but is not necessary for human survival.

wealth personal wealth is the stock of everything that a person or household owns at a particular point in time which has value.



2

Macroeconomics

The national and international economy



9 The measurement of macroeconomic performance

This is the first of the seven remaining chapters in the book covering macroeconomics. The subject area of the previous eight chapters on microeconomics describes, explains and analyses the 'little bits' of the economy: for example, individual markets, firms and industries. By contrast, macroeconomics looks at 'the economy as a whole', or in aggregate. Consider the question 'what determines the price of bread?' This is a microeconomic question, focusing on supply and demand in a single market *within* the economy. By contrast, 'what determines the average price level of *all* goods and services?' is a macroeconomic question. Similarly, 'what determines the annual rate of change of the overall price level, i.e. the rate of inflation?' is a macroeconomic rather than a microeconomic question. This and similar questions relating to the levels and rates of change of economic variables, such as output, consumption, investment, and exports and imports, lie at the heart of macroeconomics.

LEARNING OBJECTIVES

These are to understand:

- the main objectives of macroeconomic policy
- the possibility of policy trade-offs and conflicts in attempting to achieve the objectives of macroeconomic policy
- the range of data used to measure the performance of an economy
- how index numbers are calculated and interpreted, and used to show changes in key economic variables
- the uses of national income data to assess and compare living standards between countries and over time
- the importance of using purchasing power parity (PPP) exchange rates when making international comparisons.

9.1 The objectives of government economic policy

A **policy objective** is a target or goal that a government wishes to achieve or 'hit'. Since the Second World War (1939–45), governments in mixed economies such as the UK have generally had the same broad range of objectives. These are to:

KEY TERM

policy objective a target or goal that policy-makers aim to 'hit'.

- achieve economic growth and improve living standards and levels of economic welfare
- create and maintain full employment or low unemployment
- limit or control inflation, or to achieve some measure of price stability
- attain a satisfactory balance of payments, usually defined as the avoidance of an external deficit which might create an exchange rate crisis.

Economic growth

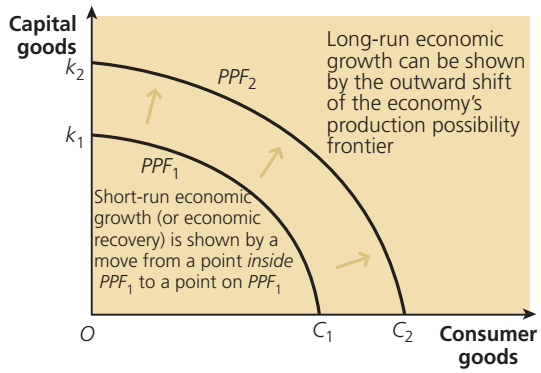


Figure 9.1 Short-run and long-run economic growth

You first came across economic growth in Chapter 1 when learning about production possibility curves. As a recap, Figure 9.1, which is much the same as Figure 1.6 in Chapter 1, will remind you of the distinction between short-run and long-run economic growth. **Short-run economic growth**, which occurs when there are unemployed resources (including labour) or 'slack' in the economy, is when there is a movement from a point *inside* the economy's production possibility frontier to a point *on* the frontier. Short-run growth is also called economic recovery. **Long-run economic growth**, by contrast, results from an outward movement of the production possibility frontier, from PPF_1 to PPF_2 in Figure 9.1.

Figure 9.2 shows what happened to the annual rate of growth of real **gross domestic product** (**real GDP**) (as distinct from **nominal GDP**) in the UK and in the **eurozone** between the first quarter of 2007 and the end of the fourth quarter of 2021. For the UK, the data for 2007 show the final year of the long economic boom that preceded the 2008–09 **recession**. Negative economic growth occurred in the recession, which ended midway through 2009, after which there was a continuous but slow recovery in the UK. The eurozone, by contrast, entered a second recession in 2012. After the Brexit referendum in June 2016, GDP growth fluctuated in both the UK and in the eurozone, but was generally less than 1% a year.

KEY TERMS

short-run economic growth growth of real output resulting from using idle resources, including labour, thereby taking up the slack in the economy.

long-run economic growth an increase in the economy's potential level of real output, and an outward movement of the economy's production possibility frontier.

gross domestic product (GDP) the sum of all goods and services, or level of output, produced in the economy over a period of time, e.g. one year.

KEY TERMS

real GDP a measure of all the goods and services produced in an economy, adjusted for price changes or inflation. The adjustment transforms changes in nominal GDP, which is measured in money terms, into a measure that reflects changes in the total output of the economy.

nominal GDP GDP measured at the current market prices, without removing the effects of inflation.

eurozone (also known as the **euro area**) the name used for the group of EU countries that have replaced their national currencies with the euro. As of 2022, 19 of the 27 EU members have replaced their national currencies with the euro.

recession in the UK and many other countries, a recession is defined as six months or more of negative economic growth or declining real national output.

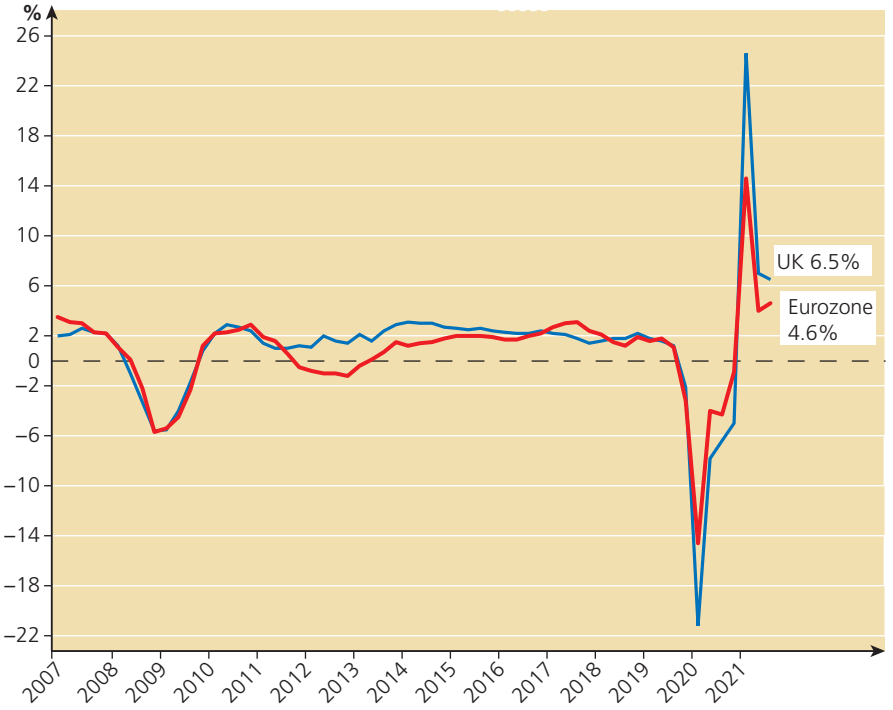
STUDY TIP

Find out what has happened to economic growth, both in the UK and in the eurozone, since the end of 2021.

Economic growth from early 2020 in both the eurozone and in the UK has been much more volatile. This is mainly the result of the Covid-19 lockdowns occurring in both 2020 and 2021 and the uncertainty that the pandemic created. The graph shows that growth fell sharply due to the enforced closure of many businesses. Ending of lockdowns led to rapid increases in GDP. However, the uncertainties over the omicron variant of Covid-19 and then the war in Ukraine in 2022 are likely to cause growth rates to be lower than expected and also more volatile.

STUDY TIP

Make sure you appreciate that economic growth is always measured in *real* rather than in *nominal* terms. You must understand the difference between real GDP and nominal GDP. (Nominal GDP is also called money GDP.) The difference between the two is that nominal GDP is real GDP multiplied by the average current price level for the year in question.



Sources: ONS, Eurostat

Figure 9.2 Year-on-year economic growth in the UK and the eurozone, quarter 1, 2007 to quarter 4, 2021

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.1

The Great Depression in the 1930s

These days, a recession is defined in the UK (though not in the USA) as a fall in real GDP which lasts for at least six months. However, a depression (or slump) is a vaguer term, best thought of as a very deep and long recession.

The 1920s was a period of growing national prosperity in the USA. Nevertheless, the Great Depression, when it arrived in 1929–30, was steeper and more protracted in the USA than in other industrial countries. The US unemployment rate rose higher and remained higher longer than in any other western country. US real GDP fell by 9.4% in 1930 and the US unemployment rate climbed from 3.2% to 8.7%. In 1931, real GDP fell by another 8.5% and unemployment rose to 15.9%. But 1932 and 1933 were the worst years of the Great Depression. By 1932, real GDP had fallen in the USA by 31% since 1929 and over 13 million Americans had lost their jobs. The US economy began the first stage of a long recovery in 1934: real GDP rose by 7.7% and unemployment fell to 21.7%.

Follow-up questions

1 What is meant by negative economic growth?
2 Find out how a recession is defined in the USA. (Search on the internet for an article by Stephen Foley, 'Economists who make the recession call', *The Independent*, January 2008.)
3 Research the details of the recession that the UK experienced in 2008–09. When did recovery from recession start, and what is the state of the economy at the time you are reading this book?
4 What is an economic recovery? Is the UK experiencing a recovery on the date you are reading this case study?



Unemployed American workers queue for free food during the Great Depression

TEST YOURSELF 9.1

Find out how many recessions there have been in the UK since 1990.

SYNOPTIC LINK

Economic growth is explained in further detail in Chapter 11 on economic performance, where we distinguish between short-run and long-run economic growth, as well as the causes of economic growth and the costs and benefits of growth for individuals, the economy and the environment.

QUANTITATIVE SKILLS 9.1

Worked example: calculating the mean and median values

Using the data in Table 9.1, calculate the mean value, to 2 decimal places, and the median value, of the annual change in gross value added (GVA) per head in the UK regions in 2019.

Table 9.1 Annual changes in gross value added per head in UK regions, 2019 (%)

Region	% change in GVA per head
London	3.8
South East	3.5
East of England	2.9
West Midlands	1.7
East Midlands	2.7
Yorkshire and Humber	3.0
South West	2.7
North West	2.7
North East	2.5
Scotland	3.0
Wales	2.1
Northern Ireland	2.0

Source: ONS, December 2021

Gross value added (GVA) is a term used in the UK's national accounts to measure the contribution to the economy of each individual producer, industry or sector in the UK. It is used in the estimation of GDP. Table 9.1 shows the percentage changes that occurred in *nominal* GVA per head for the 12 planning regions into which the UK is divided, for the year 2019 — the last year before the major disruption caused by the Covid-19 pandemic.

With a set of data such as this, the mean and the median are two measures of average values. The mean is the value obtained by dividing the sum of the percentage changes by the number of regions. Thus:

3.8% + 3.5% + 2.9% + 1.7% + 2.7% + 3.0% + 2.7% + 2.7% + 2.5% + 3.0% + 2.1% + 2.0% = 32.6%

32.6% ÷ 12 = 2.72%

So 2.72 is the mean value, to 2 decimal places.

By contrast, the median value is the 'middle number' in the sorted list of numbers. The sorted list, ranging from highest to lowest, is:

3.8% 3.5% 3.0% 3.0% 2.9% 2.7% 2.7% 2.7% 2.5% 2.1% 2.0% 1.7%

As the sorted list contains 12 numbers (one for each of the UK's 12 planning regions), the middle numbers are the sixth and seventh numbers in the list, with five numbers above and five below their value. Dividing the sum of the sixth and seventh numbers (2.7% and 2.7%) by two, the median value of 2.7% is obtained.

KEY TERM

full employment according to Beveridge's definition, full employment means 3% or less of the labour force unemployed. According to the free-market definition, it is the level of employment at the market-clearing real-wage rate, where the number of workers whom employers wish to hire equals the number of workers wanting to work.

Full employment and unemployment

There are different definitions of **full employment**, two of which we briefly explain in this chapter. First is the so-called Beveridge definition. In 1944, a famous White Paper on employment policy, written by William Beveridge (an economist at the London School of Economics, who later became Lord Beveridge), effectively committed modern governments to achieving full employment. In the White Paper, Beveridge defined full employment as occurring when unemployment falls to 3% of the labour force.

Partly because they regard Beveridge's 3% definition as too arbitrary and lacking any theoretical underpinning, free-market economists favour a second definition of full employment. For them, full employment occurs in the economy's aggregate labour market at the market-clearing real-wage rate, where the number of workers willing to work equals the number of workers whom employers wish

to hire. In Figure 9.3, this is shown where the aggregate supply curve of labour intersects the aggregate demand curve for labour. The full employment wage rate is w_{FE} and the full employment level of employment is E_{FE}.

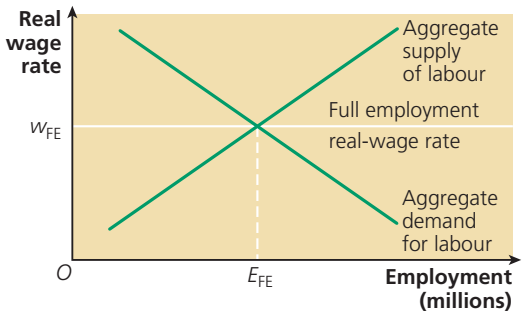
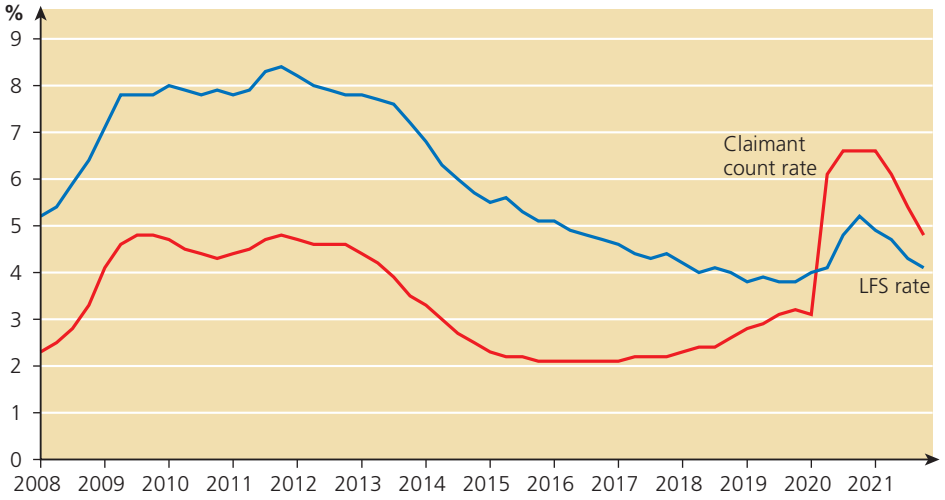


Figure 9.3 Full employment in the economy

Figure 9.3 could be interpreted as showing that when full employment occurs, there is absolutely no unemployment. However, in the real economy in which we live, this is not the case. Beveridge's definition of full employment accepts this fact. There will always be *some* unemployment, simply because the economy is constantly changing, with some jobs disappearing while new jobs are created. Chapter 11 describes and explains some of the main types of unemployment that economists recognise.

Two methods have been used to measure unemployment in the UK. One of these, the **claimant count**, has been dropped by the UK government as an official measure of unemployment because it is increasingly a poor proxy for overall unemployment levels. The second measure, the **Labour Force Survey** (LFS) measure is favoured by government and is the official measure of unemployment. As Figure 9.4 illustrates, the LFS measure has been higher than the claimant count measure. After 2013 the claimant count measure also includes recipients of Universal Credit, making the link between the two measures less comparable. This is especially true after the start of Covid-19 pandemic in early 2020 as more households relied on Universal Credit and the furlough scheme, used between April 2020 and September 2021.



Source: ONS

Figure 9.4 Changes in LFS unemployment and claimant count unemployment in the UK, quarter 1, 2008 to quarter 4, 2021

KEY TERMS

claimant count the method of measuring unemployment according to those people who are claiming unemployment-related benefits (Jobseeker's Allowance).

Labour Force Survey a quarterly sample survey of households in the UK. Its purpose is to provide information on the UK labour market. The survey seeks information on respondents' personal circumstances and their labour market status during a period of one to four weeks.

STUDY TIP

Make sure you appreciate how the LFS measure of unemployment is constructed. There are significant differences between the claimant count and LFS measures.

SYNOPTIC LINK

Employment and unemployment are explained in more detail with reference to economic performance in Chapter 11, where we consider the causes (or types) of unemployment, its consequences, and policies used to reduce the unemployment level.

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.2

UK economy: unemployment figures explained by the ONS

People often question the way the Office for National Statistics (ONS) compiles the unemployment figures. Caron Walker, director of collection and production for the ONS, explains how and why it is done:

> The answers lie with the Labour Force Survey, the huge continuous survey that ONS uses to measure unemployment (along with employment and economic inactivity). The unemployment figures use an internationally agreed definition.
>
> To count as unemployed, people have to say they are not working, are available for work and have either looked for work in the past four weeks or are waiting to start a new job they have already obtained. Someone who is out of work but doesn't meet these criteria counts as 'economically inactive'.
>
> Each quarter the LFS covers 100,000 people in 40,000 households chosen randomly by postcode. That's about one in 600 of the total population. The results are then weighted to give an estimate that reflects the entire population. Our survey has a very large sample size compared, for example, with opinion polls, which often sample around 1,000 people. Even so, in any survey there is always a margin of uncertainty, in this case around plus or minus 3% for the unemployment level.
>
> The other measure of joblessness — the claimant count — is published for each single month. It doesn't suffer from the limitations of sample size and sampling frame because it derives from the numbers of Jobseeker's Allowance (JSA) claimants recorded by Jobcentre Plus, so a monthly figure is possible right down to local level. But because many people who are out of work will not be eligible for JSA, it is a narrower measure of unemployment; it is usually much lower than the LFS measure. And it's not comparable with other countries' data, unlike the LFS measures which use internationally agreed definitions.

Follow-up questions

1. Research what has happened to the UK unemployment figures since the end of 2021.
2. How is the Labour Force Survey (LFS) measure of unemployment constructed?
3. The article by Caron Walker was published in 2013. What has happened to the UK claimant count since then?
4. Find out what has been happening to the level of unemployment in the part of the UK in which you live.

TEST YOURSELF 9.2
Who publishes the UK unemployment statistics each month?

Price stability

KEY TERM
inflation a persistent or continuing rise in the average price level.

Inflation is a general rise in average prices (a rise in the price level) across the economy. This must not be confused with a change in the price of a particular good or service within the economy. Goods' prices rise and fall all the time in a market economy, reflecting consumer choices and preferences, and changing costs. If the price of one item — say, a particular model of car — increases because demand for it is high, this is not inflation. Inflation occurs when most prices are rising by some degree across the whole economy. A change in the price of one good may, of course, lead to a change in the measured rate of inflation, particularly if spending on the item makes up a significant fraction of total consumer spending.

Achieving absolutely stable prices is not necessarily the same as controlling the rate of inflation. Absolute price stability requires a zero annual rate of inflation, with the average price level neither rising nor falling from year to year. Although a zero rate of inflation has occasionally been achieved, it is extremely rare. Much more usually, in the UK at least, controlling inflation means achieving a low inflation rate rather than absolute price stability. For most of the last two decades, successive UK governments have aimed to achieve a 2%

KEY TERMS

deflation a continuing tendency for the average price level to fall.

disinflation when the rate of inflation is falling, but still positive.

price index an index number showing the extent to which a price, or a 'basket' of prices, has changed over a month, quarter or year, in comparison with the price(s) in a base year.

consumer prices index the official measure used to calculate the rate of consumer price inflation in the UK. It calculates the average price increase of a basket of 700 different consumer goods and services.

retail prices index a measure formerly used to calculate the rate of consumer price inflation in the UK.

indexation the automatic adjustment of items such as pensions and welfare benefits to changes in the price level, through the use of a price index.

inflation rate. Until 2022, however, the inflation rate has been either a little above or a little below the 2% official target.

On occasion, notably in the economic downturn in 2009, there were fears that the inflation rate would become negative. Negative inflation, which involves a falling average price level, is called **deflation**. Make sure you don't confuse deflation with **disinflation**, which occurs when the rate of inflation is falling but is still positive. During 2022, inflation began to rise quickly, significantly above the official target. At the time of writing, the inflation rate had risen close to 8% and was expected to increase to just below 10% by autumn 2022. This is the highest UK inflation rate since 1990.

The changes that have taken place in the UK inflation rate between January 2011 and January 2022 are shown in Figure 9.5. The diagram introduces you to the fact that changes in a **price index** are used to measure the rate of consumer price inflation. Nowadays, the dominant price index is the **consumer prices index** (CPI), though before 2011 the **retail prices index** (RPI) was also used. The CPI is used by the government for setting the inflation rate target which the Bank of England tries to hit. The CPI has also been used for the **indexation** of welfare benefits, though in recent years some state benefits were frozen as part of the Conservative government's austerity policy.

The ways in which the CPI and RPI have been used to measure the rate of consumer price inflation are explained in section 9.3 of this chapter. (Note: *indices* is the plural of *index*.)

TEST YOURSELF 9.3

Explain the difference between deflation and disinflation.

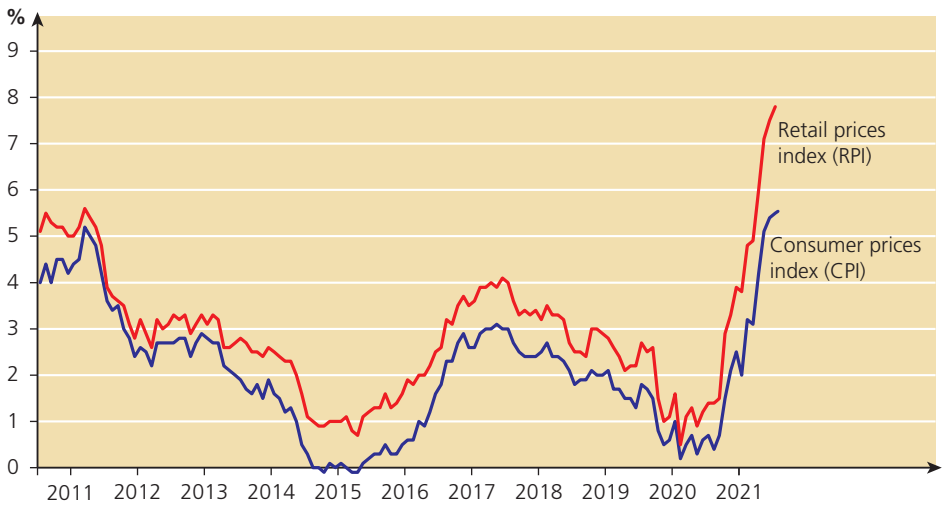


Figure 9.5 Changes in the RPI and the CPI inflation rates in the UK, January 2011–January 2022

In March 2018, the national statistician at the ONS wrote:

> Overall, RPI is a very poor measure of general inflation, at times greatly overestimating and at other times underestimating changes in prices and how these changes are experienced. In 2013, the RPI lost its status as a National Statistic. Our position on the RPI is clear: we do not think it is a good measure of inflation and discourage its use.

Until 2011, the state pension went up each year in line with changes in the RPI. The government then replaced this with the CPI link, which has generally favoured the government rather than pensioners since, as Figure 9.5 shows, RPI inflation has usually been higher than CPI inflation.

The declining usage of the RPI as a measure of inflation has accompanied the growing use of measures such as 'core' inflation, 'headline' inflation and 'underlying' inflation. 'Core' inflation is calculated by stripping out volatile items such as food and fuel from the CPI measure of inflation. 'Headline' inflation, which is currently measured by changes in the CPI, may be temporarily influenced by factors that are mainly of a short-term nature. 'Underlying' inflation, by contrast, is measured by trying to get rid of the inflationary 'noise' created by such temporary price fluctuations. In this sense, 'underlying' inflation and 'core' inflation are much the same thing.

TEST YOURSELF 9.4

Table 9.2 shows both the unemployment rate and inflation rate for an economy between 2021 and 2023.

Table 9.2 Unemployment and inflation rates, 2021–23

Year	Unemployment rate (%)	Inflation rate (%)
2021	5.4	3.3
2022	4.8	4.0
2023	4.0	4.5

What can you conclude from the data?

QUANTITATIVE SKILLS 9.2

Worked example: calculating real values from nominal values

In Ruritania between 2022 and 2023, the rate of price inflation was 6% and the rate of increase of nominal GDP was 4%. What was the rate of increase of real GDP?

To answer this question, we use the equation:

$$\text{rate of increase of real GDP} = \text{rate of increase of nominal GDP} - \text{rate of price inflation}$$

Inserting the numbers given in the question into the equation, the rate of increase of real GDP = 4% – 6%, which equals –2%.

SYNOPTIC LINK

Inflation is explained in more detail in Chapter 11, section 11.3, where we consider the causes of inflation, its consequences and the policies used to control the rate of inflation.

STUDY TIP

You don't need detailed technical knowledge of the construction of price indices such as the CPI, but you should appreciate the underlying features of a price index (see section 9.3).

QUANTITATIVE SKILLS 9.3

Worked example: calculating the rate of real economic growth

Based on the data in Table 9.3, calculate the rate of real economic growth between 2022 and 2023.

Table 9.3 Rate of real economic growth between 2022 and 2023

Year	Nominal GDP ($ billion)	Price level
2022	1,270	100
2023	1,410	108

Economic growth requires us to know the annual change in real GDP, so we need to calculate the level of real GDP In 2023.

$$\text{real GDP in 2023} = \text{nominal GDP in 2023} \times \frac{\text{price level in 2022}}{\text{price level in 2023}}$$

$$= 1{,}410 \times \frac{100}{108} = \$1{,}305.6\text{bn}$$

$$\text{economic growth in 2023} = \frac{\$1{,}305.6 - \$1{,}270}{\$1{,}270} \times 100 = 2.8\%$$

TEST YOURSELF 9.5

Distinguish between the nominal value and the real value of an economic variable.

A satisfactory balance of payments

The **balance of payments** measures the flows of money into and out of an economy crossing national boundaries in a particular time period, usually a month, quarter or year. An important part of the balance of payments is called the current account.

The **current account** of the balance of payments contains two main sections: the money value of **exports**, and the money value of **imports** (of both goods and services). Changes in these over a period extending from 2010 to 2021 are shown in Figure 9.6. Taken together, the money value of exports and imports make up the **balance of trade**. If the money value of imports exceeds the money value of exports, there is a **balance of trade deficit**; if the money value of imports is less than the money value of exports, there is a **balance of trade surplus**. Achieving balance on the current account of the balance of payments is one of the macroeconomic objectives.

KEY TERMS

balance of payments a record of all the currency flows into and out of a country in a particular time period.

current account measures all the currency flows into and out of a country in a particular time period in payment for exports and imports of goods and services, together with primary and secondary income flows (previously known as 'income flows and transfers').

exports domestically produced goods or services sold to residents of other countries.

imports goods or services produced in other countries and sold to residents of this country.

KEY TERMS

balance of trade the difference between the money value of a country's imports and its exports. Balance of trade is the largest component of a country's balance of payments on current account.

balance of trade deficit when the money value of a country's imports exceeds the money value of its exports.

balance of trade surplus when the money value of a country's exports exceeds the money value of its imports

STUDY TIP

Knowledge of the four sections of the current account of the balance of payments will be required — both for solving calculation problems and for attempting to analyse the causes of a surplus or deficit on this part of the balance of payments.

SYNOPTIC LINK

The current account of the balance of payments is explained in further detail in Chapter 14, section 14.3, where we consider the various accounts contained within the balance of payments, policies used to correct imbalances on the current account, as well as the significance of current account deficits and surpluses on the balance of payments.

TEST YOURSELF 9.6

Explain the difference between the balance of trade and the balance of payments on current account.

There are some other sections of the current account of the balance of payments, namely primary and secondary income flows. These are explained in Chapter 14.

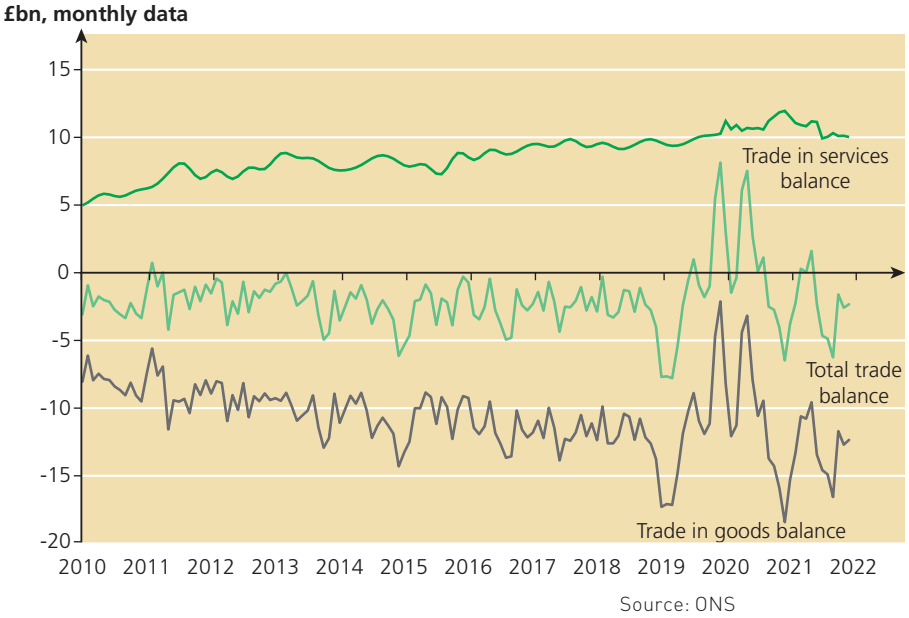


Figure 9.6 Changes in the UK's balances of trade in goods and services, and total trade, 2010–21

The word 'satisfactory' can be interpreted in different ways. People sometimes assume that a satisfactory balance of payments occurs only when the government achieves the biggest possible current account surplus (i.e. when the value of exports exceeds the value of imports by the greatest amount). However, a country can enjoy a trading surplus only if at least one other country suffers a trading deficit. It is mathematically impossible for all countries to have a current account surplus at the same time. Therefore, most economists take the view that a 'satisfactory' balance of payments is a situation in which the current account is in equilibrium, or when there is a small surplus or a small but sustainable deficit.

Other macroeconomic objectives

KEY TERMS

balanced budget achieved when government spending equals government revenue ($G = T$).

budget deficit occurs when government spending exceeds government revenue ($G > T$).

Governments may also have other objectives of macroeconomic policy, such as **balancing the budget** — the difference between government spending and tax revenue. After the 2008–09 recession, the **budget deficit** (where government spending exceeds revenue from tax) increased significantly. The reduction of this deficit became a very important macroeconomic objective for the UK government until 2015 when, although the deficit remained (albeit, a smaller deficit), other priorities emerged.

The deficit rose again to very high levels in 2021. This was the direct result of the massive increase in government spending related to the Covid-19 pandemic (both in terms of economic support, such as the furlough scheme, as well as increases in health and social care spending) and the sharp fall in taxation revenue as incomes and spending fell significantly. However, the reduction of this deficit was not a priority, as it had been ten years earlier. This is explained in Chapter 13 on fiscal policy and supply-side policies.

Another macroeconomic objective is achieving a more equitable or fairer distribution of income. In recent years the opposite has been the case: after the 2008–09 recession, income inequalities widened, although they were actually slightly narrower by 2021.

STUDY TIP

Make sure you are aware that governments have policy objectives other than achieving economic growth, low unemployment, a low inflation rate and a satisfactory balance of payments.

SYNOPTIC LINK

Chapter 1, section 1.1, introduced you to the difference between positive and normative statements in economics. Chapter 7, section 7.1, explained how unequal and inequitable distributions of income illustrate the difference between positive and normative statements. At this stage, it would be useful to reread the relevant sections in both chapters, as these will show you that governments often have to make decisions which may be partly political rather than solely economic.

Macroeconomic policy conflicts

Economists often argue that it is difficult, if not impossible, for a government to 'hit' all its desired macroeconomic objectives at the same time. Believing they can't achieve the impossible, policy-makers often settle for the lesser goal of 'trading off' between policy objectives. A **trade-off** exists when two or more desirable objectives are mutually exclusive. As the government thinks it cannot achieve, for example, full employment and zero inflation, it aims for less than full employment combined with an acceptably low and sustainable rate of inflation.

Over the years UK macroeconomic policy has been influenced and constrained by four significant conflicts between policy objectives. The main **policy conflicts** and their associated policy trade-offs are:

- between the internal policy objectives of **full employment** and **growth** and the external objective of achieving a satisfactory **balance of payments** (or possibly supporting a particular **exchange rate**)
- between achieving **full employment** and **controlling inflation**
- between increasing the rate of **economic growth** and achieving a **more equal distribution of income and wealth**
- between **higher living standards now** and **higher living standards in the future**. This is an example of an 'intertemporal' conflict (a conflict 'across time').

KEY TERMS

trade-off between policy objectives although it may be impossible to achieve two desirable objectives at the same time (e.g. zero inflation and full employment), policy-makers may be able to choose an acceptable combination lying between the extremes, e.g. 2% inflation and 4% unemployment.

policy conflict when two policy objectives cannot both be achieved at the same time: the better the performance in achieving one objective, the worse the performance in achieving the other.

SYNOPTIC LINK

Policy conflicts are explained in more detail in Chapter 11, section 11.4, in regard to economic performance.

Most economists agree that these policy conflicts and trade-offs pose considerable problems for governments in the economic short run, defined as a period in macroeconomics extending just a few years into the future. However, there is much less agreement about whether they need be significant in the long run — a period extending many years into the future. Pro-free-market economists often argue that if appropriate (and successful) supply-side policies are implemented, the main objectives of macroeconomic policy are compatible with each other and not in conflict in the long run. (Supply-side policies are explained in depth in Chapter 13.)

STUDY TIP

In Part 1 of this book, on microeconomics, you learnt that the short run is the time period in which at least one factor of production is held fixed, while the long run is the time period in which all factors of production are variable. However, in macroeconomics, economists tend generally to use the terms in a looser way. For example, the macroeconomic 'short run' may extend to about three years into the future, with the 'long run' being any period longer than that. To complicate things still further, a period known as the 'medium term' is sometimes identified. This could be a period of about 18 months to three years into the future, separating the short run and the long run.

This matters as different groups of economists have different beliefs about the exact timescales of the short-run and long-run periods. If an economic problem occurs that economists assume will resolve itself in the long run, some economists may think this is justification for not attempting to solve the problem at all (if the long run is viewed as not too far into the future). Other economists may be more willing to intervene and solve problems if they think the long run might be a significantly long time away.

SYNOPTIC LINK

In Chapter 10, a very famous quote from the economist John Maynard Keynes appears, which seeks to explain why governments should intervene in the macroeconomy rather than wait until the long run, when the problem may resolve itself.

EXTENSION MATERIAL

How the importance attached to the different macroeconomic policy objectives has changed over time

The order in which the four main objectives of macroeconomic policy were listed at the beginning of this chapter shows a broadly Keynesian ranking of priorities. In the Keynesian era, which extended roughly from the 1950s to 1979, UK governments implemented Keynesian macroeconomic policies. They believed that economic policy should be used to achieve full employment, economic growth and a generally acceptable or fair distribution of income and wealth. These were the prime policy objectives, which had to be achieved in order to increase

KEY TERM

Keynesian economists followers of the economist John Maynard Keynes, who generally believe that governments should manage the economy, particularly through the use of fiscal policy.

human happiness and economic welfare — the ultimate policy objective. Controlling inflation and achieving a satisfactory balance of payments were regarded as intermediate objectives, or possibly as constraints, in the sense that an unsatisfactory performance in controlling inflation or the balance of payments could prevent the attainment of full employment and economic growth.

In the early 1980s, things changed. A new government in 1979, with Margaret Thatcher becoming prime minister, meant that UK governments were now **pro-free market** rather than Keynesian. In the 1970s inflation had threatened to escalate out of control, and in response UK governments placed control of inflation in the top position as a policy objective, relegating full employment to a lower position in the ranking of macroeconomic policy objectives.

Since then, UK governments have continued to give much more attention to the need to control inflation. Indeed, in 1993, the Conservative chancellor Norman Lamont stated that high unemployment was a 'price well worth paying' for keeping inflation under control. This view was echoed in 1998 when, under a Labour government, the governor of the Bank of England argued that 'job losses in the north were an acceptable price to pay for curbing inflation in the south'. These statements reflect the pro-free-market view that, in order to maintain a high and sustainable level of employment, inflation must first be brought under control.

The long and deep recession which hit the UK (and many other countries) in 2008 led to a partial revision of this view. This was reinforced by the lockdown-induced downturns of 2020 and 2021 (which technically may not actually have been recessions). For various reasons, which are explained in Chapter 11, section 11.3, inflation was thought to have been successfully controlled and was kept close to its target rate for much of the period between 2012 and 2021. The rapid rise in the inflation rate in 2022 was considered to have been caused by factors outside the government's control, namely rapid rises in energy prices, oil price increases and shortages and disruptions to supply chains caused by the Covid-19 pandemic (though others blamed Brexit and the war in Ukraine for some of these disruptions). Chapters 12 and 13 explain how a combination of 'loose' **monetary policy** (i.e. very low interest rates) and 'tight' **fiscal policy** (i.e. cutting government spending in an effort to reduce the size of the budget deficit) have dominated recent UK macroeconomic policy.

It is also worth noting that, despite a rapid increase in the UK's balance of payments deficit on current account, achieving a **balance of payments equilibrium** has not been viewed by recent governments as an important policy objective. All this may, of course, change.

KEY TERMS

pro-free-market economists opponents of Keynesian economists, who dislike government intervention in the economy and who much prefer the operation of free markets.

monetary policy the use by the government and its agent, the Bank of England, of interest rates and other monetary instruments to try to achieve the government's policy objectives.

fiscal policy the use by the government of government spending and taxation to try to achieve the government's policy objectives.

balance of payments equilibrium a situation in which a deficit or surplus on the current account of the balance of payments is exactly matched by capital inflows or outflows in the other parts of the balance of payments.

STUDY TIP

During your course of study, you should try to build a good knowledge of developments in the UK economy and government policies over the 15 years or so before you started studying economics. Knowledge of earlier economic history may also be useful.

EXTENSION MATERIAL

The growth of Keynesian economics

Although the A-level specification only requires you to know about trends and developments in the economy which have taken place over the 15 years before you sit the examination, in order to understand the nature of modern macroeconomics, it is useful to go back to what was happening in the UK and world economies in the 1920s and 1930s. Most economists at the time, especially those in UK and American universities, were free-market economists who believed that, in a competitive market economy, market forces would automatically deliver full employment and economic growth. In so far as governments needed to have a macroeconomic policy, it was generally believed that the policy should be restricted to maintaining the 'sound money' deemed necessary for a stable price level, and possibly to maintaining a fixed exchange rate.

However, the problem was that in the UK economy of the 1920s, and in the wider world economy (especially the USA) in the 1930s, free-market forces did *not* deliver full employment and economic growth. Instead, unregulated market forces seemed to have produced economic stagnation and mass unemployment. The most seminal event of the time was the Great Depression, which began in 1929 and lasted until the late 1930s. Unemployment rose in 1933 to almost 25% in the USA, and in 1931 to 24% in the UK. Regional unemployment in towns such as Jarrow in northeast England rose as high as 70%, though London, the southeast and the Midlands fared much better.

Free-market economists responded to the Great Depression by arguing that markets were not to blame for persistent large-scale unemployment. Instead, they believed that mass unemployment was caused by institutional factors, such as the power exercised by trade unions, which prevented markets from operating freely. In the free-market view, wage cuts were necessary to 'price the unemployed into jobs', but trade unions resisting wage cuts prevented this from happening.

In the late 1920s, John Maynard Keynes, who started his academic career as an economist in the free-market tradition, began to change his view on the main cause of unemployment. In response to an accusation of inconsistency, Keynes is reported to have said: 'When the facts change, I change my mind — what do you do, sir?'

Keynes believed that orthodox free-market economic theory failed to explain how the whole economy works, and that a better and more general theory was needed to explain mass unemployment. Keynes completed his new theory in 1936 with the publication of his great book *The General Theory of Employment, Interest and Money*, commonly referred to as Keynes's *General Theory*.

Keynes's *General Theory* marks the beginning of modern macroeconomics. For over a generation until about 1979, Keynesian economics was macroeconomics, and macroeconomics was Keynesian economics. In the three decades after the Second World War, Keynesianism became the new economic orthodoxy in the UK, the Netherlands and the Scandinavian countries. Economic policy in the USA also eventually became Keynesian, with the Republican president, Richard Nixon, famously stating in 1971 that 'we are all Keynesians now'.



Jarrow marchers protest against unemployment and poverty in 1936

SECTION 9.1 SUMMARY

- A policy objective is a target or goal that a government wishes to achieve.
- Over the years, the main policy objectives of UK governments have been to achieve economic growth and improve living standards and levels of economic welfare, to create and maintain full employment or low unemployment, to limit or control inflation, and to attain a satisfactory balance of payments.
- Governments may also have other objectives of macroeconomic policy, such as balancing the budget and achieving a more equitable or fairer distribution of income.
- Short-run growth or economic recovery is a movement from a point *inside* the economy's production possibility frontier to a point *on* the frontier.
- Long-run growth results from an outward movement of the production possibility frontier.
- The Beveridge definition of full employment relates to unemployment falling to 3% of the labour force.
- Free-market economists define full employment as occurring at the market-clearing real-wage rate, where the number of workers willing to work equals the number of workers whom employers wish to hire.
- Inflation is a general rise in average prices (a rise in the price level) across the economy.
- In the UK, controlling inflation means achieving a low inflation rate rather than absolute price stability.
- A satisfactory balance of payments is a situation in which the current account is in equilibrium, or when there is a small surplus or a small but sustainable deficit.
- Over the years, UK macroeconomic policy has been influenced and constrained by four significant conflicts between policy objectives: between achieving full employment/growth and achieving a satisfactory balance of payments (or a particular exchange rate); between achieving full employment and controlling inflation; between increasing the rate of economic growth and achieving a more equal distribution of income and wealth; and between higher living standards now and higher living standards in the future.

9.2 Macroeconomic indicators

KEY TERM
macroeconomic indicators provide information from recent economic performance for judging the success or failure of a particular type of government policy (e.g. fiscal policy or monetary policy).

The A-level specification defines **macroeconomic indicators** as data which are commonly used to measure the performance of an economy, such as real GDP, real GDP per capita, the consumer prices index, measures of unemployment, productivity and the balance of payments on current account.

Macroeconomic indicators provide governments and their policy-makers with information about the recent success or lack of success of a particular type of economic policy, such as monetary policy or fiscal policy. The size and rate of change of the money supply is a commonly used monetary performance indicator. Likewise, the size and rate of change of the budget deficit is an indicator of fiscal policy performance.

Macroeconomic indicators also provide information about whether current policy is on course to hit the future target set for the stated policy. For many years, a 2% rate of inflation, measured by the annual change in the consumer prices index, has been a policy indicator for the success of monetary policy. Likewise, achieving a balanced budget and seeing the national debt falling as a percentage of GDP are targets used to measure the success of fiscal policy in the medium term. An indicator, such as information about labour productivity and productivity gaps, can also be used to compare the performance of the UK economy with that of competitor countries.

Macroeconomic indicators can be divided into *lead* and *lag* indicators:

- **Lead indicators** provide information about the future state of the economy (stemming from the way people are currently forming their expectations).

Surveys of consumer and business confidence and investment intentions indicate the existence of a feel-good or feel-bad factor and provide information about the likely state of aggregate demand a few months ahead. Statistics for house-building starts and the number of people who have booked a holiday several months in advance also provide information about future spending, while data on commodity and input prices can signal future changes in retail price inflation.

- **Lag indicators** provide information about past and possibly current economic performance and the extent to which policy objectives such as economic growth and control of inflation have been achieved. Data on the level of GDP, and current and recent employment and unemployment figures, provide examples of lag indicators giving information about current and recent economic performance.

Macroeconomic indicators are almost always presented in the form of statistical data: for example, unemployment and growth figures in the case of lagged indicators, and projections about the number of house-building starts in the case of lead indicators. The accuracy of the information provided by performance indicators is thus highly dependent on the accuracy of the statistics available from the government and other sources.

TEST YOURSELF 9.7

Suppose there is an increase in people booking overseas holidays for the summer ahead. Would this be a lead indicator or a lag indicator? Explain your answer.

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.3

David Smith's skip index and other confidence indicators

Every week David Smith, the economics editor of the *Sunday Times*, writes an Economic Outlook column — highlights can be accessed on David Smith's **economicsuk.com** website.

A few years ago, David Smith came up with the idea of a 'skip index', as an informal lead indicator of what might happen to the economy in the future. The idea here is that people use builders' skips when undertaking extensive building work, such as a loft conversion or an extension to a property. This would indicate that either people feel wealthy or they are confident that they can profit from improving property. However, the accuracy of a skip index can be questioned. While an increase in the number of builders' skips might mean people are more affluent and spending the money on their houses, it might also mean that people cannot afford to move but their family is still growing and therefore they have to upgrade their house. A 'scaffolding index' might suffer from the same problem as a skip index. Of course, this is perhaps less relevant if fewer people can afford to buy their own houses, anyway.

Another possibility is a 'crane index'. This is a way to gauge prosperity by counting the cranes on the urban skyline. However, it is also possible to turn a crane index on its head. An increase in the number of tower cranes may indicate over-expansion or over-confidence — not necessarily economic growth. Another confidence indicator is frequency of receipt of unsolicited letters from estate agents, claiming for example that a 'Mr Jones' is desperate to buy a house in your road if you'll just give the agent a call.

Follow-up questions

1. Construct a skip index for your local community. What conclusions can you draw from the information you collect?
2. Can you think of any other things that could be used to gauge the state of consumer or business confidence in the economy?
3. Questionnaires and surveys are sometimes used to find out what people think will happen to the economy in the future. Why should the information provided by such surveys be treated with caution?
4. Inspired by David Smith's skip index, a trade group, the Building Merchants Federation, now publishes a Builders Merchant Building Index (BMBI). Find out what the latest quarterly BMBI report has to stay about confidence in the building industry.

SECTION 9.2 SUMMARY

- Macroeconomic indicators are data which are commonly used to measure the performance of an economy, such as real GDP.
- Macroeconomic indicators provide governments with information about the recent success of a particular type of economic policy, such as monetary policy or fiscal policy.
- Macroeconomic indicators also provide information about whether current policy is on course to hit the future target set for the stated policy.
- An indicator can also be used to compare the performance of the UK economy with that of competitor countries.
- Macroeconomic indicators can be divided into *lead* and *lag* indicators.
- Lead indicators provide information about the future state of the economy, stemming from the way people are currently forming their expectations.
- Surveys of consumer and business confidence and investment intentions provide information about the likely state of aggregate demand a few months ahead.
- Lag indicators provide information about past and possibly current economic performance and the extent to which policy objectives such as economic growth and control of inflation have been achieved.
- Data on the level of GDP and current and recent employment and unemployment figures are examples of lag indicators.
- Macroeconomic indicators are almost always presented in the form of statistical data: for example, unemployment and growth figures in the case of lag indicators, and projections about the number of house-building starts in the case of lead indicators.
- The accuracy of the information provided by performance indicators is thus highly dependent on the accuracy of the statistics available from the government and other sources.

9.3 The uses of index numbers

Introduction to index numbers

Earlier in this chapter, we mentioned the difference between nominal GDP (or money GDP) and real GDP. To recap, real GDP, or real gross domestic product, is a measure of the total quantity of goods and services produced by the economy over a period of time, having got rid of the distortive effects of price changes or inflation.

KEY TERM
index number a number used in an index (e.g. the consumer prices index) to enable accurate comparisons over time to be made.

Changes in real GDP, along with other economic variables, are usually expressed using **index numbers**. (Earlier in the chapter we explained how the consumer prices index (CPI) has largely replaced the retail prices index (RPI) as a measure of the price level and the rate of inflation.) Because index numbers frequently appear in the quantitative data you are expected to interpret in the course of your studies, it is especially important that you build up an understanding of how economic indices are constructed.

Economists frequently use index numbers when making comparisons over periods of time. An index starts in a given year, called the base year, which is usually given an index number of 100. In later years or months, an increase in the size of the variable causes the index number to rise above 100, while a fall in the size of the variable, compared to the base year, results in the index number falling to below 100. For example, an index number of 105 means a 5% rise from the base year, whereas an index number of 95 means a 5% fall.

QUANTITATIVE SKILLS 9.4

Objective test question: understanding data presented in index numbers, GDP per head

Table 9.4 shows index numbers for GDP per head for two countries, L and M, in 2015 and 2020.

Table 9.4 Index numbers for GDP per head, 2015 and 2020

	GDP per head	
	2015	2020
Country L	100	115
Country M	100	108

Which one of the following statements is supported by the data?

A GDP per head was the same in both countries in 2015.
B GDP rose faster in country L than in country M.
C The rate of increase in GDP per head was greater in country L than in country M.
D GDP per head was higher in country L than in country M in 2020.

To answer the question correctly, it is important to appreciate that the data show GDP *per head of population*, and not the overall GDP figures for both countries. This means that, although statement B could well be true, it cannot be concluded from the data, as the rate of population growth could be slower in country M than in country L. Statements A and D are wrong for a different reason, namely that there is insufficient information in the data to allow us to compare *across* the two data series. For statement A, the index number is the same in both countries, namely 100. The base year for both data series is probably 2015. The fact that both index numbers are 100 tells us nothing about whether GDP per head was the same in both countries in this year. For the same reason, we cannot conclude that GDP per head was higher in country L in 2020, simply because 115 is a larger number than 108. This means statement D is wrong. However, while we cannot compare *across* the two data series, we can compare *along* each series taken in isolation. The data allow us to conclude that GDP per head increased by 15% in country L between 2015 and 2020, but by only 8% in country M. This means that statement C is the correct answer.

QUANTITATIVE SKILLS 9.5

Objective test question: interpreting macroeconomic data

Table 9.5 shows figures for population and index numbers for inflation (CPI) and money national income (GDP at current prices) in the years 2022 and 2023.

In 2023, compared to 2022, which one of the following statements can be inferred from the data?

A Population grew at a slower rate than prices.
B The inflation price index increased by 25%.
C Money GDP per head rose by 6%.
D Real GDP increased.

Statement A provides the correct answer. The price level increased by nearly 67% ((50/75) × 100), but population grew by only 5%. Statement B is wrong because the starting-off price index was 75 and not 100. Statement C is wrong because, although money GDP did indeed increase by 6%, the population grew in size by 5%, so the increase in money GDP *per head* was less than 6%. Finally, statement D is also wrong: if money GDP increased by 6% but the price level increased by nearly 67%, *real* GDP must have fallen.

Table 9.5 Population size and GDP and inflation indices, 2022 and 2023

Year	GDP at current prices	Inflation index	Population size
2022	100	75	40 million
2023	106	125	42 million

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.4

The construction of the consumer prices index (CPI)

The following passages have been extracted from an article, 'Consumer price indices, a brief guide: 2017', published by the Office for National Statistics (ONS) in 2017 and accessible on the ONS website.

This article describes the main consumer price indices in the UK and explains how they are put together. It focuses on the consumer prices index (CPI) including owner occupiers' housing costs (CPIH) but briefly mentions the retail prices index (RPI).

Consumer price indices

Consumer price indices are important indicators of how the UK economy is performing. The indices are used in many ways by the government, businesses and society in general. They can affect interest rates, tax allowances, wages, state benefits, pensions, maintenance, contracts and many other payments. They also show the impact of inflation on family budgets. The CPIH and CPI measure price changes, not price levels. They are therefore expressed in terms of the comparison of prices relative to 2015, when the indices were given values of 100.

Inflation figures

The CPIH is the most comprehensive measure of inflation. It extends the CPI to include a measure of the costs associated with owning, maintaining and living in one's own home, known as owner occupiers' housing costs, along with Council Tax. Both of these are significant expenses for many households that are excluded from the CPI.

The RPI — a long-standing measure of UK inflation — does not meet the required standard for designation as National Statistics. The RPI, its subcomponents and the RPIX (the retail prices index excluding mortgage interest costs) continue to be published as they are tied to long-term contracts, related for example to wage and saving contracts.

The shopping basket

A convenient way of thinking about the CPIH and CPI is to imagine a very large 'shopping basket' full of goods and services on which people typically spend their money: from bread to ready-made meals, from the cost of a cinema seat to the price of a pint at the local pub, from a holiday in Spain to the cost of a bicycle. The content of the basket is fixed for a period of 12 months.

The cost of living

The CPIH and CPI are not 'cost of living' indices. The index simply indicates what we would need to spend in order to purchase the same things we bought in an earlier period, irrespective of whether particular products are 'needed' or are 'good for you'.

What's in the basket?

The CPIH and CPI 'baskets' include all types of household spending. It is impractical and unnecessary to monitor the price of every product sold in every single shop. The prices of similar items can reasonably be assumed to move in line with one another in response to market forces. It is therefore sufficient to compile the index using prices of a large and varied sample of products in selected locations. The goods and services for which prices are recorded are called 'representative items'.

The CPIH, CPI and RPI are fixed quantity price indices: specific items are chosen to represent price movements in the baskets; prices are collected for those items only (700 representative items). Their movements are taken to represent the price changes for all goods and services covered by the index, including those for which prices are not specifically monitored.

Quality changes

It is important that the index calculations are based on like-for-like comparisons, of prices each month for each of the items in the basket. However, some brands or varieties of particular products priced at the start of the year may not be available in later months. Adjustments are also made, for example when a manufacturer changes the size or weight of a product.

Weighting

The quantities or 'weight' of the various items in the basket are chosen to reflect their importance in the typical household budget. We spend more on some things than others. So we would expect, for example, a 10% increase in the price of petrol to have a much bigger impact on the CPIH or CPI than a similar rise in the price of tea. For this reason, the components of the index are 'weighted' to ensure that it reflects the importance of the various items in the average shopping basket, and the amounts we spend in different regions of the country and in different types of shops.

Updating the shopping basket

It is important that the index is representative and kept up to date. The basket of goods and services is therefore reviewed every year, helping to ensure that the CPIH and CPI calculations more accurately reflect UK shopping and purchasing patterns.

Some changes to the basket are necessary each year due to changing markets, fashions and new products. Smart-phones and tablet PCs, for example, have been added in recent years. The basket is fixed for a period of 12 months.

[Note: the basket was also updated in 2022 due to Covid-19-related factors as anti-bacterial wipes were added to the basket. Pet collars was also included to represent the increased spending on pets — a probable response to more people spending time at home during lockdowns. Doughnuts were removed from the basket to reflect decreased spending on this item.]

The weights for the index are also changed each year to keep pace with general changes in our spending habits. Over the years people have tended to spend more of their money on electrical goods, travel and leisure while the proportion they spend on basics has fallen.

[Note: it is likely that the weights attached to the categories covering domestic heating and transport costs will rise significantly to reflect the sharp increase in heating bills and petrol prices that started in late 2021 and continued into 2022.]

Calculating the index

After the price data have been checked and processed, the resulting price indicators are combined. Changes in prices of the individual goods and services in the index are measured by comparing them to their levels in the previous January. These are then weighted together using the weights for the current year to produce an overall average price change. The final stage in the calculation is to link the average price changes with the figures for earlier years. Only by 'chain-linking' the calculations in this way can the index take account of changes in the make-up of the shopping basket from year to year and provide like-for-like comparisons between different years — this procedure ensures that the index is not distorted when items are either removed from, or introduced into, the CPIH or CPI 'shopping basket'.

Follow-up questions

1 Distinguish between the price level and the rate of inflation.
2 What is meant by 'weighting' in the construction of a price index such as the CPI?
3 Why may changes in a price index such as the CPI not provide an accurate measure of the rate of inflation for an individual: for example, a pensioner?
4 Why has the consumer prices index (CPI) largely replaced the retail prices index (RPI) as a measure of the average price level?



The consumer prices index calculates the average price increase of a 'basket' of 700 goods and services

STUDY TIP

Although a number of different price indices are produced, the only two that are on the A-level specification are the CPI and RPI.

TEST YOURSELF 9.8

What does a price index attempt to measure?

QUANTITATIVE SKILLS 9.6

Worked example: calculating a weighted price index

Table 9.6 shows details for a weighted price index containing three items: X, Y and Z.

Table 9.6 Weighted price index for three products

Product	Year 1 price (£)	Year 2 price (£)	Weight
X	6	9	1
Y	12	18	3
Z	20	15	2

If year 1 is the base year, what is the value of the weighted price index in year 2 and what would be the inflation rate in year 2?

Based on year 2 prices, the year 2 price index for each item would be as follows:

product X: (9 ÷ 6) × 100 = 150; product Y: (18 ÷ 12) × 100 = 150; product Z: (15 ÷ 20) × 100 = 75

If we multiply each item in the index by its weight, then for year 2 we have:

X: 150 × 1 = 150; Y: 150 × 3 = 450; Z: 75 × 2 = 150 giving a total of 150 + 450 + 150 = 750

The total of the price index divided the total of the weights (1 + 3 + 2 = 6) is 750/6 = 125

The inflation rate then is 25%.

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.5

Spending by the average UK family

The goods and services in the CPI shopping basket and the weights attached to them are meant to reflect the spending patterns of the 'average British family'. In so far as there is such a thing, the ONS calculated in 2021 that in 2020 the average British household spent £587.90 a week in the financial year ending (FYE) 2020 (see Figure 9.7). This was slightly down in real terms when compared with the previous year (spending in the FYE 2019 was £603.10 per week).

By looking at headline economic indicators, we can put the latest household spending estimates into a wider context. Inflation, as measured by the CPI, including owner occupiers' housing costs (CPIH), fell from 2% in April 2019 to 1.5% in March 2020, the lowest level of inflation seen since late 2016.

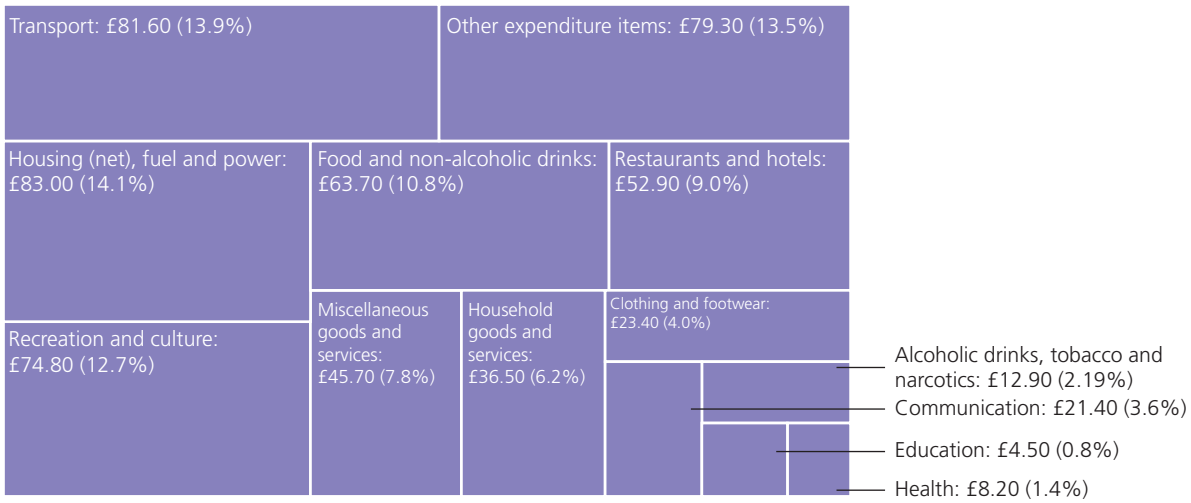
GDP, a headline measure of how well the economy is performing, had grown steadily for all quarters since quarter 1, 2013 (January to March). However, although 2020 probably started off with positive growth, the final figure for quarter 1, 2020 showed a decline in GDP — the result of the slowdown in spending and output caused by the start of the Covid-19 pandemic, culminating with the lockdown in the last part of this quarter. Employment also continued to grow during this period. The employment rate reached 76.6% in January 2020; at that time, the highest rate since records began. The higher the number of people and households there are in work, the more households there are receiving a regular wage; this could be another factor that influenced higher levels of spending in this time period. However, there was a slight fall in the employment rate by March 2020 (down to 76% — and this continued to fall in 2020) which was probably the result of the slowdown in spending and uncertainty created at the start of the pandemic.

Median disposable income increased by £1,214 (or 4.1%) when compared with the previous year. This shows that households had more money to spend when compared with the previous year.

Alongside these main economic indicators, there are other indicators we can consider to help build a wider picture of the economy. For most of 2019, the household savings ratio fell to between 4 and 5%, the lowest seen since 1999. This shows that households chose not to save their money during this period, which could be attributed to the low base interest rate (Bank Rate) for most of the FYE 2020 of 0.75% — which, although it had remained at this level since 2018, was still low by historic standards. This rate was reduced twice in March 2020 to 0.1% by the end of the FYE 2020. A low interest rate is a disincentive to save but it is an incentive to spend. In particular, it can encourage households to buy more goods and services on credit, as well as taking out loans, which would influence higher levels of household spending. However, the lockdown and uncertainty of the pandemic led to the savings ratio rising in quarter 1, 2020 to 6.7% (and it rose significantly after this period).

Follow-up questions

1 Using examples taken from Figure 9.7, explain how higher levels of employment affect household spending.
2 What is the household savings ratio? How does a falling savings ratio affect household spending?
3 Figure 9.7 shows family spending on education in the financial year ending in 2020. Why do you think family spending on education was so low?



Source: ONS

Figure 9.7 How the 'average family' spent £587.90 a week in the financial year ending 2020

EXTENSION MATERIAL

Other economic indices

In addition to the RPI, CPI and CPIH, introduced earlier, the Office for National Statistics also publishes a number of price indices which are not directly related to changes in consumer prices. These include producer price indices (PPI), which measure changes in input and output prices of goods bought and sold by UK manufacturers. Input prices are prices of materials and fuels bought; output prices, also known as 'factory gate prices', are prices at which goods are sold.

Other indices constructed by the ONS measure changes in a range of economic variables including output per worker or labour productivity, mentioned earlier in this chapter. The ONS also publishes the index of production, which measures the volume of production at base year prices for the manufacturing, mining and quarrying, and energy supply industries.

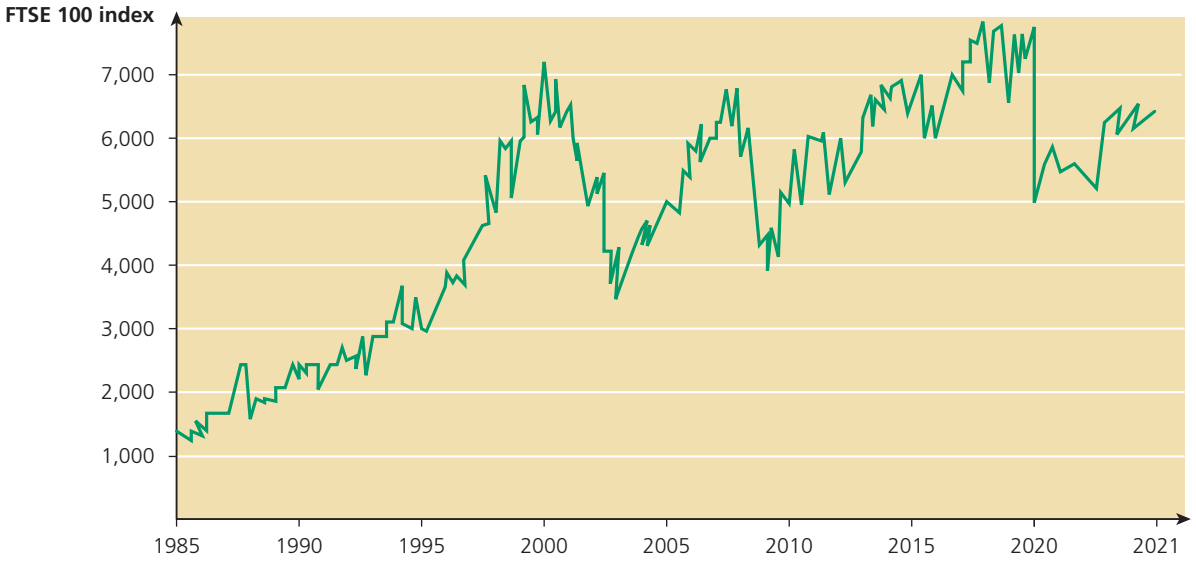
EXTENSION MATERIAL

The 'Footsie' 100 index

Economic and financial indices are also published by non-government organisations. Perhaps the best known in the UK is the FTSE 100 index, often called the 'Footsie' index. The FTSE 100 is an index composed of the 100 largest companies listed on the London Stock Exchange (LSE). The index is seen as a good indicator of the performance of major public companies (PLCs). The FTSE Group which produces the index is 50/50 owned by the *Financial Times* and the London Stock Exchange (hence FTSE — FT and SE). Although the FTSE 100 is the most famous index the company produces, the FTSE Group also calculates other indices, covering markets around the world, every day. In the UK, the other FTSE indices include the FTSE 250 (the next 250 largest companies after the FTSE 100) and the FTSE All-Share index.

One point to note about the FTSE 100 index is that it was launched in 1984 with a base year value of 1,000. Usually, as we have seen, the base year index number in an economic index is 100. However, as the FTSE 100 index shows, this does not have to be the case. Also, unlike most other indices, including those published by the ONS, the FTSE 100 index has never been rebased. (Rebasing means that the base year used in an index is changed every few years, with the index number 100 being assigned to the new base year.) In early 2020, the FTSE 100 index was very close to reaching 8000 before it fell significantly as the economic implications of the Covid-19 pandemic became understood. The fact that the FTSE index has never been rebased allows us to see the index as a barometer of UK economic performance over the last three decades.

Figure 9.8 shows how the FTSE 100 index has risen (and at times fallen) over the years between 1984 and 2021. However, the graph does not show how the composition of the UK's 100 leading companies has changed over that period. Companies are promoted into and relegated from the FTSE index every three months. Back in 1984, the index was dominated by the UK's leading manufacturing companies. This is no longer the case. Leading manufacturers have either disappeared or been taken over by companies based overseas. To figure in the FTSE index, a company's shares have to be quoted on the London Stock Exchange. When a UK company is taken over and becomes a subsidiary of an overseas-based company, it disappears from the FTSE index. These days, the FTSE 100 companies are largely energy companies, retail companies and other service industry companies.



Source: OpenLearn

Figure 9.8 Changes in the FTSE 100 index, 1984–2021

SECTION 9.3 SUMMARY

- Changes in real GDP, and in other economic variables, are usually expressed using index numbers.
- An index number is a number used in an index, such as the consumer prices index, to enable accurate comparisons over time to be made.
- In 2017 household spending exceeded pre-economic downturn levels for the first time.
- GDP fell rapidly due to the uncertainty and forced closure of many businesses by the government as a result of the Covid-19 pandemic and lockdowns.
- Rising levels of inflation can encourage households to bring forward certain purchases.
- This in turn boosts economic growth.
- The number of people and households that are in work could be another factor that influences levels of spending.
- In 2019 the savings ratio fell below 5% — the lowest for over a decade — before it rose as a response to the pandemic.
- A low interest rate is a disincentive to save but an incentive to spend.

9.4 The uses of national income data

National income

KEY TERMS

national income the *flow* of new output produced by the economy in a particular period (e.g. a year), measured by the flow of factor incomes.

national product (also known as **national output**) the *flow* of new output produced by different industries in a particular period (e.g. a year).

national capital stock the stock of capital goods (e.g. buildings and machinery) in the economy that has accumulated over time and is measured at a point in time.

human capital the skills, knowledge and experience possessed by the population.

national wealth the stock of all goods that exist at a point in time that have value in the economy.

STUDY TIP

Make sure you understand the difference between stocks and flows. Stocks measure variables at a point in time, and flows measure variables over a period of time.

Economists use the terms **national income** (NI) and **national product** (NP), which is also called **national output**, interchangeably. To produce the *flow* of national income or output (i.e. the total income or output earned or produced over a period of time, rather than at a point in time), the economy must possess a *stock* of physical capital goods (the **national capital stock**) and a stock of **human capital**, together with stocks of the other factors of production: land and entrepreneurship. The national capital stock is part of the stock of **national wealth**, which comprises all physical assets owned by the nation's residents that have value. The national capital stock includes capital goods, together with social capital such as the roads, hospitals and schools which are owned by the state. However, the national capital stock excludes consumer goods, which are a part of national wealth but not part of national capital. All capital is wealth, but not all wealth is capital.

STUDY TIP

Make sure you don't confuse *wealth*, which is a stock at a point in time, with *income*, which is a flow generated over a period of time. Wealth is the stock of assets, or things that have value, which people own.

Figure 9.9 shows how we can relate a country's national income or output (e.g. in 2023) to the national capital stock and the human capital stock. In the figure, national income in 2023 and 2024 is shown as the area contained by the three rectangles *A*, *B* and *C*. We will now assume that at the beginning of 2023, the economy operates on its production possibility frontier. This means there is no unemployed labour, and the economy is working at full capacity to produce the flow of national income shown by $A + B + C$.

The national capital stock is the stock of capital goods that has accumulated over time in the economy. By contrast, national income is the flow of new output produced by the stocks of physical and human capital and other factors of production.

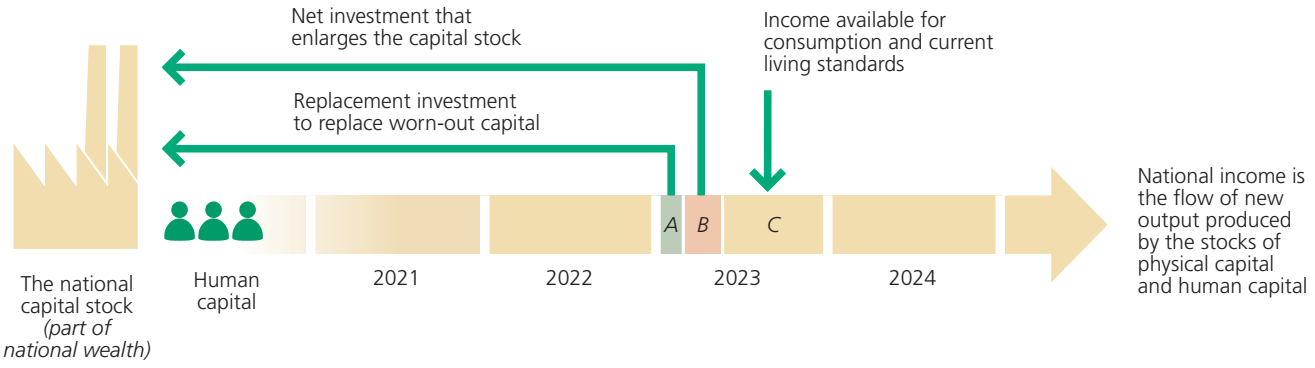


Figure 9.9 National income and the stocks of physical and human capital

TEST YOURSELF 9.9
Distinguish between positive and negative economic growth and explain how negative economic growth relates to recessions.

However, part of the national capital stock wears out in the course of producing 2023's national income. Worn-out capital (or the *depreciation* of the national capital stock) reflects *capital consumption*. To maintain the size of the capital stock, so that (in the absence of population growth and technical progress) the stock is capable of producing exactly the same size of national income in 2024 as in 2023, part of 2024's national output must be replacement capital goods. In other words, some of the investment spending in an economy will be replacing worn-out capital rather than adding to the capital stock of investment goods. The spending on replacement investment is shown by area *A*. If this investment doesn't take place, the national capital stock shrinks. Negative economic growth occurs and the economy's production possibility frontier shifts inwards.

Positive economic growth generally requires that investment takes place, over and above the replacement investment shown by rectangle *A* in Figure 9.9. The extra investment needed to enlarge the capital stock is shown by rectangle *B*. This is called *net investment*. Gross investment (shown by *A* + *B*) is the sum of replacement investment and net investment. Only net investment increases the size of the capital stock, thereby facilitating long-run (positive) economic growth.

KEY TERM
consumption total planned spending by households on consumer goods and services produced within the economy.

Rectangle *C* shows the fraction of national income available for **consumption** in 2023. A decision to sacrifice current consumption in favour of a higher level of future consumption means that more of society's scarce resources go into investment or the production of capital goods, enabling the national capital stock to increase in size. However, in the short term, the easiest way to increase living standards is to boost current consumption. This 'live now, pay later' approach sacrifices saving and investment, which ultimately reduces long-term economic growth.

TEST YOURSELF 9.10
What would happen to economic growth if gross investment were negative?

EXTENSION MATERIAL

The difference between national income and GDP

Students often confuse gross national income (GNI) with gross domestic product (GDP), though GDP is now the most commonly used measure of national income. Both reflect the national output and income of the economy. The word 'domestic' in gross domestic product indicates that GDP is the flow of output produced domestically within the economy. However, some UK residents and companies receive income from assets they own in other countries. Likewise, some of the income produced within the UK flows out of the country to overseas owners of assets located within the UK. GNI and gross national product (GNP) take account of these income flows into and out of the economy; GDP does not.

As it measures all the income available to spend in the UK, GNI is a better measure of the standard of living currently enjoyed by UK residents. However, GDP is a better measure of the productivity of industries actually located within the UK. Over the years, the UK invested in and accumulated capital assets in other countries. As a result, in many years income flows into the UK were larger than income flows leaving the UK. This meant that UK GNI and GNP were larger than UK GDP. Since 2011, the reverse has been true — UK GDP has been larger than GNI and GNP. Thus in 2020 GNI was £2,121,181 million, but GDP was larger at £2,152,646 million.

STUDY TIP
Although useful to know, GNI is not on the AQA specification. It is only GDP (real and nominal) that is used as the measure of national income.

The use and limitations of national income data to assess changes in living standards *over time*

Because GNP or GNI, GDP and other national income statistics are the main source of data on what has happened and what is happening in the economy, they are often used as indicators of economic growth, economic and social welfare and changing living standards, and for comparison with other countries.

To see how living standards change over time, we must look at figures for real per capita GNP, which is real GNP divided by the number of people living in the country. Rising real GNP per capita gives a general indication that living standards are rising, but it may of course conceal great and sometimes growing disparities in income distribution. This is especially significant in economically developing countries where the income distribution is typically extremely unequal and where only a small fraction of the population may benefit materially from economic growth.

Besides the problem of income distribution, a number of other problems surface when using national income statistics to measure living standards. The problems include:

- **The non-monetised economy.** In the UK, national income statistics under-estimate the true level of economic activity because housework and 'do-it-yourself' home improvement take place without money incomes being generated. When measuring national income, a decision has to be made about whether to estimate or to ignore the value of this production. The UK accounts can be criticised for estimating the value of some but not all of the non-monetised economy. 'Imputed rents' are estimated for the 'housing services' delivered to owner-occupiers by the houses they live in, based on an estimate of the rent that would be paid if the house-owners were tenants of the same properties. But house-keeping allowances paid within households are not estimated, implying that housework is unproductive. Judgements such as these lead to the anomaly that national income appears to fall when, for instance, a man stops paying for a house cleaner and does the housework himself.
- **The hidden economy.** Economic activity undertaken illegally in the **hidden economy** may be omitted. The hidden economy refers to all the economic transactions conducted in cash which are not recorded in the national income figures because of tax evasion. By its nature, it is impossible to make a completely accurate estimate of the size of the hidden economy. The gap between the GNP total obtained by the income and expenditure methods of measurement (see page 329) can be used to approximate the size of the hidden economy. It probably equals about 10% of the UK's measured GNP, while some other countries, such as Greece and Italy, have hidden economies equal to 20–30% of GNP.

KEY TERM
hidden economy (also known as the **informal economy**, the **underground economy** and the **black economy**) all the economic transactions conducted in cash which are not recorded in the national income figures because of tax evasion.

- **Quality changes.** Over time, the quality of goods changes for better or worse, presenting a particularly difficult problem in the construction and interpretation of national income figures. This is true also of services. When services such as public transport and health deteriorate, GNP may rise even though welfare and real living standards decline.
- **Negative externalities.** National income statistics over-estimate living standards because of the effects of negative externalities such as pollution and congestion, and activities such as crime. What is in effect a welfare loss may be shown as an increase in national output, falsely indicating an apparent welfare gain. For example, the stresses and strains of producing an ever-higher national output lead to loss of leisure time, and people become ill more often. Loss of leisure and poorer health cause welfare to fall. But in the national accounts, these show up as extra production and as extra consumption of healthcare, both of which imply a welfare gain. Traffic congestion increases the cost of motoring, and hence the value of national income. Motorists would prefer uncongested roads and less spending on petrol and vehicle wear and tear.

The use and limitations of national income data to compare differences in living standards *between countries*

National income data also have limitations when comparing living standards in different countries:

- **Relative importance of the non-monetised economy.** Comparisons of national income per head between countries are misleading if the relative importance of the non-monetised economy differs significantly. There are also differences in the degree of statistical sophistication in data collection, particularly between developed and developing economies, and a lack of international uniformity in methods of classifying and categorising the national accounts. There are further problems in making comparisons when different commodities are consumed. Expenditure on fuel, clothing and building materials is usually greater in economically developed countries located in parts of the world with cold winters than in much warmer developing economies. However, greater expenditure on these goods may not indicate higher real incomes and living standards.
- **Exchange rates.** It is possible to compare GNP per capita in different countries by converting GNP figures for each country into a common currency such as the US dollar. However, such calculations suffer from the assumption that the exchange rates between local currencies and the dollar are correctly valued, in the sense that a dollar's worth of output in one country becomes immediately and accurately comparable with a dollar's worth of output in any other country. This can never be so. Exchange rates can only correctly reflect the values of internationally traded goods such as automobiles or internationally traded foodstuffs and raw materials.
- **Traded and non-traded goods.** The purchasing power of a currency over domestically produced goods and services that do not enter into international trade or compete domestically with imports may be completely different from the currency's purchasing power over imported goods. Exchange rate changes only reflect the price changes of internationally traded goods. In so far as there is a much wider gap in developing

TEST YOURSELF 9.11

Why may changes in market exchange rates lead to misleading comparisons of the standard of living in different countries?

economies than in developed economies between the price changes of internationally traded and non-traded goods, GNP figures measured in US dollars tend to under-estimate real levels of income and output in developing economies.

Purchasing power parity (PPP) exchange rates and comparing living standards

KEY TERM
purchasing power parity (PPP) exchange rates the rates of currency conversion that equalise the purchasing power of different currencies by eliminating the differences in price levels between countries.

The correct solution to the problem outlined in the previous section is to establish **purchasing power parity (PPP) exchange rates**.

Calculations of GDP based on market exchange rates tend to over-estimate the cost of living in poorer developing countries. PPP exchange rates are based on the idea that, in the long run, exchange rates should move towards rates that equalise the prices of an identical basket of goods and services in any two countries. Stated simply, a dollar, or any other unit of currency such as the euro, should buy the same everywhere. Purchasing power is determined by the relative cost of living and inflation rates in different countries. Purchasing power parity means equalising the purchasing power of two currencies by taking into account these cost of living and inflation differences.

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.6

Market exchange rate versus purchasing power parity (PPP) exchange rates: which measurement should we use?

In everyday life it is usual to use a market exchange rate, the rate prevailing in the foreign exchange market, to calculate, for example, what a UK pound is worth in terms of US dollars. In some circumstances however, economists prefer to use PPP exchange rates — the rate at which the currency of one country would have to be converted into that of another country to buy the same amount of goods and services in each country.

In an article titled 'PPP versus the market', published in the March 2007 edition of *Finance and Development*, senior IMF official Tim Callen wrote:

> To understand PPP, let's take a commonly used example, the price of a hamburger. If a hamburger is selling in London for £2 and in New York for $4, this would imply a PPP exchange rate of 1 pound to 2 US dollars. This PPP exchange rate may well be different from that prevailing in financial markets (so that the actual dollar cost of a hamburger in London may be either more or less than the $4 it sells for in New York). This type of cross-country comparison is the basis for the well-known 'Big Mac' index, which is published by the *Economist* magazine and calculates PPP exchange rates based on the McDonald's sandwich that sells in nearly identical form in many countries around the world.

Of course, any meaningful comparison of prices across countries must consider a wide range of goods and services. This is not an easy task, because of the amount of data that must be collected and the complexities in the comparison process. To facilitate price comparisons across countries, the International Comparisons Program (ICP) was established by the United Nations and the University of Pennsylvania in 1968. PPPs generated by the ICP are based on a global survey of prices.*

So which method is better? The appropriate way to aggregate economic data across countries depends on the issue being considered. Market exchange rates are the logical choice when financial flows are involved. For example, the current account balance represents a flow of financial resources across countries. It is appropriate to use the market exchange rate to convert these flows into dollars when aggregating across regions or calculating the global current account discrepancy. But for other variables, the decision is less clear cut. Take real GDP growth. International organizations use different approaches. The World Bank uses market-based rates to determine the weights in its regional and global aggregations of real GDP, whereas the IMF and the Organization for Economic Cooperation and Development use weights based on PPP. Each methodology has its advantages and disadvantages.

Advantages of PPP. A main one is that PPP exchange rates are relatively stable over time. By contrast, market rates are more volatile, and using them could produce quite large swings in aggregate measures of growth even when growth rates in individual countries are stable.

Another drawback of market-based rates is that they are relevant only for internationally traded goods. Non-traded goods and services tend to be cheaper in low-income than in high-income countries. A haircut in New York is more expensive than in Lima; the price of a taxi ride of the same distance is higher in Paris than in Tunis; and a ticket to a cricket game costs more in London than in Lahore. Indeed, because wages tend to be lower in poorer countries, and services are often relatively labour-intensive, the price of a haircut in Lima is likely to be cheaper than in New York even when the cost of making tradable goods, such as machinery, is the same in both countries.

Any analysis that fails to take into account these differences in the prices of non-traded goods across countries will underestimate the purchasing power of consumers in emerging market and developing economies and, consequently, their overall welfare. For this reason, PPP is generally regarded as a better measure of overall well-being.

Drawbacks of PPP. The biggest one is that PPP is harder to measure than market-based rates. It is a huge statistical undertaking, and new price comparisons are available only at infrequent intervals. The method of measurement does not cover all countries, which means that data for missing countries have to be estimated.

*Note that an update for 2020 was produced and estimates were released covering recent years. Each of the participating countries — 176, in total — provides national average prices for 1,000 closely specified products.

Follow-up questions

1 Access the OECD website at **https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm** and identify the PPP exchange rates for the UK pound and the US dollar for the latest year on the chart.
2 Why do you think that the US dollar has the same PPP exchange rate in every year on the chart?
3 Find out from the internet or from the business page of a recent newspaper the market exchange rate of the UK pound against the US dollar.
4 What conclusion(s) can you draw from comparing your answers for the two exchange rates?

APPLICATION OF ECONOMICS IN THE REAL WORLD 9.7

The Economist's Big Mac index

In 1986 *The Economist* magazine invented the Big Mac index as a light-hearted guide to whether currencies are at their 'correct' level. The Big Mac index is based on the theory of purchasing power parity (PPP), the notion that in the long run exchange rates should move towards the rate that would equalise the prices of an identical basket of goods and services (in this case, a Big Mac burger) in any two countries. 'Burgernomics' was never intended as a precise gauge of currency misalignment, merely as a tool to make exchange-rate theory more digestible. The GDP-adjusted Big Mac index addresses the criticism that you would expect average burger prices to be cheaper in poor countries than in rich ones because labour costs are lower. PPP signals where exchange rates should be heading in the long run, as a country like China gets richer.

Follow-up questions

1 Distinguish between a market exchange rate and a PPP exchange rate.
2 Why is a Big Mac burger a particularly suitable commodity for measuring a PPP exchange rate?
3 Name another commodity which might also be suitable for measuring a PPP exchange rate and explain why.
4 Find out from the Big Mac index whether the UK pound's exchange rate is currently over-valued or under-valued.



One way of measuring purchasing power parity

SYNOPTIC LINK

Exchange rates are explained in Chapter 15, section 15.1, where we look at the determinants of the exchange rate, as well as how government policy matters in deciding what will happen to the value of a country's currency.

SECTION 9.4 SUMMARY

- National income data are especially important for monitoring the performance of the UK economy.
- Economists use the terms national income (NI) and national product (NP) interchangeably.
- Gross domestic product (GDP) is now the most commonly used measure of national income.
- Because GNP or GNI, GDP and other national income statistics are the main source of data on what has happened and what is happening in the economy, they are often used as indicators of economic growth, economic and social welfare, and changing living standards, and for comparison with other countries.
- National income statistics under-estimate the true level of economic activity because the non-monetised economy is under-represented.
- The hidden economy refers to all the economic transactions conducted in cash which are not recorded in the national income figures because of tax evasion.
- Standards of living reflect elements of quality of life and general economic welfare, as well as narrow GNP-related consumption.
- Comparisons of national income per head between countries are misleading if the relative importance of the non-monetised economy differs significantly.
- Purchasing power parity (PPP) exchange rates are the rates of currency conversion that equalise the purchasing power of different currencies by eliminating the differences in price levels between countries.

Questions

1 With the help of a *PPF* diagram, explain the difference between short-run and long-run economic growth.

2 Distinguish between an objective and an instrument of macroeconomic policy. Explain two macroeconomic policy objectives.

3 Explain two of the main conflicts affecting macroeconomic policy, and how a government might trade off between achieving these objectives.

4 Explain the differences between inflation, deflation and disinflation. When did deflation last occur in the UK economy?

5 What is the difference between a deficit and a surplus on the current account of the balance of payments? Which does the UK have today, and how large is it?

6 Briefly explain how a price index is constructed.

7 What is the purpose of a PPP exchange rate?



10

How the macroeconomy works: the circular flow of income, aggregate demand/aggregate supply analysis and related concepts

This chapter is about *macroeconomic theory*. The two main bodies of macroeconomic theory, or macroeconomic models, are explained in some detail. These are the circular flow and the aggregate demand/aggregate supply (*AD*/*AS*) models of the economy. The chapter then goes on to look at related macroeconomic concepts, such as the components of aggregate demand (consumption, investment etc.) and the multiplier concept. The following three chapters of the book, which cover economic performance, monetary policy and fiscal policy, will continue the task introduced in this chapter of applying macroeconomic theory to the analysis of current economic problems and the evaluation of government economic policies.

LEARNING OBJECTIVES

These are to remind you of the meaning of national income, and then to understand:

- what national income measures and the difference between nominal and real income
- the circular flow of income and the concept of equilibrium national income
- the determinants of the aggregate demand (*AD*) curve
- the determinants of and differences between the short-run (*SRAS*), long-run (*LRAS*) and the Keynesian aggregate supply curves
- the multiplier process and how it is affected by the marginal propensity to consume
- the analysis of *AD* and *AS* and how their interaction affects the price level and the level of real GDP.

10.1 The circular flow of income

National income revisited

In order to understand the meaning of the circular flow of income, it is useful to remind yourself of what we wrote about national income in the previous chapter. In Chapter 9, we explained:

- **What national income measures.** National income is the flow of new output produced by an economy in a particular time period, such as a year.
- **The difference between nominal and real national income.** The flow of new output is real national income, which comprises the actual goods and services that are produced. By contrast, nominal national income measures the flow of output at the current price level in the economy.
- **The role of real national income as an indicator of economic performance.** The *level* of real national income is an indicator of current living standards within the economy, and its *rate of change* measures the economic growth (or economic decline) occurring in the economy. The level and rate of change of real national income are used when comparing the country's economic performance with that of other countries.

National income = national output = national expenditure

It is also important to understand that there are three methods of measuring the flow of new output in an economy. First, the flow of new output can be measured by adding together the incomes received by the different factors of production such as capital and labour. Measured in this way, the flow is called *national income*. Secondly, the flow can be measured by summing the totals of the actual goods and services produced by the economy. The flow is now called *national output* or *national product*. Finally, we can measure *national expenditure*. This shows how factor incomes such as wages and profits end up being spent on the goods and services produced by the economy.

Since they are three different ways of measuring the same thing (the flow of new output), and provided that we ignore income flows into and out of the economy, it follows that national income, national output and national expenditure always equal each other and add to the same totals.

In summary:

- The **income approach** sums the factor incomes to the factors of production.
- The **output approach** adds up the 'value added' by each of the industries in the economy, such as agriculture, manufacturing and services, by including the increase in the value of output at each stage of production (i.e. how much extra is added on at each intermediate stage in the production process).
- The **expenditure approach** sums consumption + investment + government expenditure + exports – imports.

The preliminary estimate of gross domestic product (GDP) is based solely on the output approach to measuring the flow of new output. The output approach measures gross value added (GVA) at a detailed industry level before aggregating to produce an estimate for the whole economy.

STUDY TIP

It might be useful at this stage to reread section 9.4 in Chapter 9 on the uses of national income data to re-familiarise yourself with how national income is calculated, as well as the uses of the statistic as an economic indicator (and the limitations on its usefulness).

TEST YOURSELF 10.1
Distinguish between real and nominal national income and explain why real national income is a better indicator of economic performance than nominal national income.

The circular flow of income

Figure 9.9 in Chapter 9 depicts national income as the flow of new output moving along a horizontal arrow extending from the economy's national capital stock. Look back also to Figure 6.1 in Chapter 6. The diagram shows households and firms functioning simultaneously in both the goods market, or market for output, and the economy's factor market or market for inputs into the production process. However, the roles of households and firms are reversed in these two sets of markets. Whereas firms are the source of supply in a goods market, in a factor market firms exercise demand for factor services supplied by households.

We noted in Chapter 6 that the incomes received by households from selling their factor services largely enable households to demand the output supplied by firms in the goods market. Demand requires both a willingness and an ability to pay, so households need an income, and for most people this means they have to sell their labour services in a labour market.

The relationship between households and firms in the two markets is therefore circular. In goods markets, finished goods and services flow from firms to households, who spend their incomes on the goods. In labour markets, members of households sell their labour to firms in order to earn the incomes that they spend on goods and services.

This is the basis of the circular flow of income which we illustrate in the next three diagrams.

KEY TERM
closed economy an economy with no international trade.

The simple circular flow model, illustrated in Figure 10.1, assumes that there are just two sets of economic agents in the economy: households and firms. The model initially assumes that the government and foreign trade do not exist (these assumptions will be relaxed later in this section). This is a model of a two-sector economy, or a **closed economy** with no government sector. The dashed flow lines in the figure show the real flows occurring in the economy between households and firms. Households supply labour and other factor services in exchange for goods and services produced by the firms. These real flows generate money flows of income and expenditure, shown by the solid flow lines.

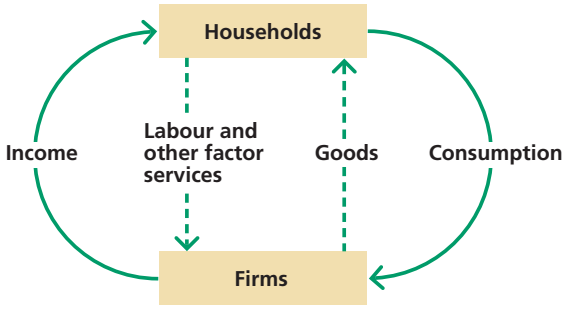


Figure 10.1 A simplified circular flow diagram of a two-sector economy

KEY TERMS

saving income which is not spent.

withdrawal a leakage of spending power out of the circular flow of income into savings, taxation or imports.

investment total planned spending by firms on capital goods produced within the economy.

injection spending entering the circular flow of income as a result of investment, government spending and exports.

All the income received by households (shown by the left-hand flow curve of the diagram) is spent on consumption (shown by the right-hand flow curve).

Figure 10.2 is a more realistic version of Figure 10.1, because it shows households saving as well as consuming, and firms investing in capital goods. When households save part of their incomes, people are spending less than their incomes. **Saving**, which is an example of a leakage or **withdrawal** from the circular flow of income, is depicted by the upper of the two horizontal arrows in Figure 10.2. The lower of the two horizontal arrows shows **investment**, or spending on machinery and other capital goods, which is an **injection** of demand into the economy.

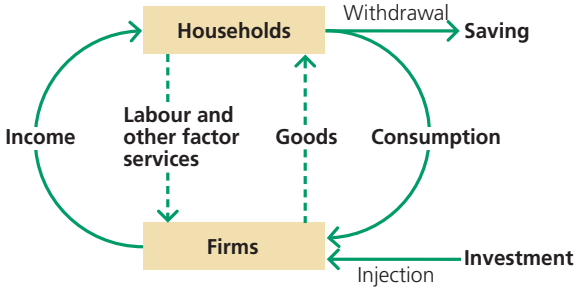


Figure 10.2 Introducing saving and investment into the circular flow of income

A first look at the concept of equilibrium national income

Equilibrium national income is the level of income at which withdrawals from the circular flow of income equal injections into the flow. And as we will shortly see when taking a second look at the concept, it is also the level of output at which aggregate demand equals aggregate supply. Equilibrium income is not necessarily the same as **full employment income**. This is the level of income when the economy is producing on its production possibility frontier, with no spare capacity.

TEST YOURSELF 10.2

Using examples, explain the difference between an injection into, and a withdrawal from, the circular flow of income.

KEY TERMS

equilibrium national income the level of income at which withdrawals from the circular flow of income equal injections into the flow; also the level of output at which aggregate demand equals aggregate supply.

full employment income the level of income when the economy is producing on its production possibility frontier, with no spare capacity.

How a change in injections or withdrawals affects equilibrium national income

In circular flow theory, if planned saving (or the planned withdrawal of spending) equals planned investment (or the planned injection of spending into the flow), national income is in equilibrium, tending neither to rise nor to fall. However, if the withdrawal exceeds the injection, the resulting net leakage of spending from the circular flow causes output and income to fall. Likewise, if the withdrawal is less than the injection, the resulting net injection of spending into the economy causes output and income to rise.

TEST YOURSELF 10.3

Name two other contexts in which economists use the term 'equilibrium' (this can be an equilibrium from either microeconomics or macroeconomics).

Savings can be hoarded (i.e. stored away), or the funds being saved can be lent for others to spend. Hoarding — for example, keeping money under the mattress, but more likely saving it in a bank account — means that a fraction of income is not spent. This can lead to deficient aggregate demand in the economy, which means there is too little demand to buy the output the economy is capable of producing. As they cannot sell some of the goods and services they have produced, firms reduce their output and national income falls.

However, if all savings are lent, via financial intermediaries such as banks, for firms and other consumers to borrow and then to spend, planned saving may end up equalling planned investment. With this outcome, national income remains in equilibrium and there is no reason why the level of income should fall.

In the two-sector circular flow model, national income is in equilibrium when:

planned saving = planned investment

or:

$S = I$

TEST YOURSELF 10.4
If people hoard much of their incomes, what adverse (i.e. harmful) effects can occur in the economy?

However, because households and firms have different motives for making their respective saving and investment decisions, there is no reason why, initially, household saving should exactly equal the amount firms plan to spend on capital goods (i.e. investment). Consider, for example, a situation in which planned saving is greater than planned investment ($S > I$). In this situation, the national income or output circulating round the economy is in disequilibrium, with withdrawals out of the system exceeding injections of spending into the flow of income, and national income will fall.

Bringing the government sector and the overseas sector into the circular flow model

The simple circular flow model we have looked at is perhaps simplistic because it ignores how the domestic government and the rest of the world are sources of both injections of demand into, and withdrawals of demand from, the circular flow of income. This may seem unrealistic — governments are usually significant contributors to total spending in an economy and very few economies do not engage in significant amounts of foreign trade.

Government spending is an injection of spending into the economy. In contrast, taxation, which takes spending power away from the people being taxed, is a withdrawal or leakage of spending. Exports and imports result from the fact that we live in an **open economy**: that is, an economy open to international trade. Exports, which result from people living in other countries buying goods produced in the UK, are an injection into the circular flow. By contrast, imports, or spending by UK residents on goods produced overseas, are a withdrawal from the circular flow. Spending on imports stimulates the economies of other countries rather than the UK economy.

KEY TERM
open economy an economy open to international trade.

Figure 10.3 shows the effect of bringing the activities of the government and overseas sectors into the circular flow model.

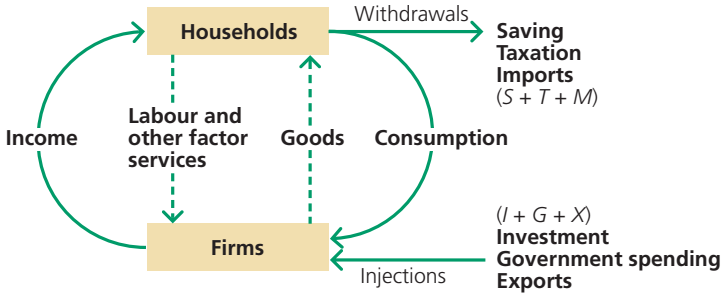


Figure 10.3 The circular flow of income in an economy with a government sector and an overseas sector

In this extended circular flow model, national income is in equilibrium, tending neither to rise nor to fall, when:

saving + taxation + imports = investment + government spending + exports

or:

$$S + T + M = I + G + X$$

However, whenever:

$$S + T + M > I + G + X$$

a net withdrawal or leakage of demand out of the circular flow occurs, which in Keynesian analysis causes the equilibrium level of national income to fall.

And when:

$$S + T + M < I + G + X$$

a net injection of demand into the circular flow occurs, which causes the equilibrium level of national income to rise.

STUDY TIP

Economists use the letters *Y*, *C*, *S*, *I*, *G*, *T*, *X* and *M* as shorthand, respectively, for the macroeconomic variables income, consumption, saving, investment, government spending, taxation, exports and imports. In the exam, make sure you don't confuse investment and income — we use *I* for investment and *Y* for income. Also, be careful: *S* refers to savings here but can also be used as an abbreviation of Supply.

EXTENSION MATERIAL

Keynes and deficient aggregate demand

If planned saving by households exceeds planned investment by firms, there is a danger that deficient aggregate demand may cause the economy to sink into a recession. In the 1930s, during the Great Depression, John Maynard Keynes argued that if household savings are not lent to finance spending by others, particularly investment by firms, the level of income or output circulating round the economy falls. This reduces saving, until planned saving equals planned investment and equilibrium is restored, but at a significantly lower level of national income. The economy ends up in recession. In recent years, Keynesian economists used the same analysis as part of their explanation of the 2008/09 recession, arguing that a liquidity crisis and an associated collapse of consumer and business confidence were responsible for the collapse in aggregate demand.

Pro-free-market economists, who generally reject Keynesian theory, believe that deficient aggregate demand only exists as a *temporary* and self-correcting phenomenon. They argue that, when deficient aggregate demand occurs, the rate of interest falls, rather than the level of income or output, quickly restoring equality between saving and investment intentions. When interest rates fall, people save less because saving becomes less attractive. At the same time, firms invest more in new capital goods because the cost of borrowing has fallen.

Back in the 1930s, Keynes agreed that a fall in interest rates can bring about equality between saving and investment, but he believed the process to be slow. In the very long run it may work, but, in Keynes's memorable phrase, 'in the long run we are all dead'. Keynes argued that, when planned leakages of demand from the circular flow of income exceed planned injections of demand into the flow, the level of income or output falls to restore equilibrium. According to Keynes, deficient aggregate demand is the cause of recessions.

EXTENSION MATERIAL

Building and using economic models

Much of this chapter is about macroeconomic models: the circular flow model we have just explained, and the *AD/AS* model we are about to investigate, are the two macroeconomic models you need to understand. Model building is one of the most fundamental analytical techniques used by economists. Economic theory is based on developing economic models which describe particular aspects of the economic behaviour of individuals, groups of individuals and, in a macro context, the whole economy.

A model is a small-scale replica of real-world phenomena, often incorporating a number of simplifications. An economic model simplifies the real world in such a way that the essential features of an economic relationship or set of relationships are explained using diagrams, words and often algebra. Models are used by economists first to understand and explain the working of the economy, and secondly to predict what might happen in the future.

A good economic model simplifies reality sufficiently to allow important and often otherwise obscure economic relationships to be studied, away from irrelevant detail or 'background noise'. The danger is that reality can be over-simplified, with the resulting model failing to reflect in a useful way the world it seeks to explain. Economic modelling involves the art of making strong assumptions about human behaviour so as to concentrate attention and analysis on key economic relationships in a clear way, while avoiding an excessive over-simplification of the problem or relationship to be explained.

The ultimate purpose of model building is to derive predictions about economic behaviour, such as the prediction of demand theory that demand will increase when price falls. Economic controversy often exists when models generate conflicting predictions about what will happen in a given set of circumstances. For example, a model of the labour market which predicts that the supply of labour increases as wages rise carries the policy-making implication that a cut in income tax, being equivalent to a wage rise, creates an incentive to work more. Under alternative assumptions, the model could predict the opposite: that, as wages rise, workers begin to prefer leisure to work and react to the tax cut by working less.

It may often be possible to accept or dismiss a model of economic behaviour on the basis of common sense or casual observation of the world around us. Economists now usually go further, using sophisticated statistical tests to evaluate empirically the model's predictions. Good economic models or theories survive the process of empirical testing (which is part of a branch of the subject called 'econometrics'), whereas models or theories shown to be at odds with observed behaviour must be revised or discarded.

TEST YOURSELF 10.5
How is a recession defined in the United Kingdom?

STUDY TIP
Many macroeconomic questions can be addressed by using the *AD/AS* model (and a smaller number, the circular flow model).

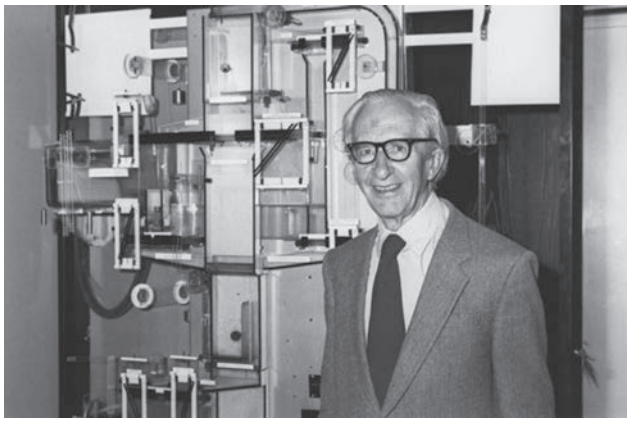
SYNOPTIC LINK
Measures of economic performance in Chapter 11, and different economic policies discussed in Chapters 12 and 13, are explained and analysed using one or both of the macroeconomic models (circular flow and *AD/AS*) explained in this chapter.

EXTENSION MATERIAL

Circular flow and the Phillips machine

In the Meade Room in the Department of Applied Economics of Cambridge University stands a Phillips machine — a device so cunning and ingenious that it can predict the running of the national economy to within 4% accuracy. The prototype was an odd assortment of tanks, pipes, sluices and valves, with water pumped around the machine by a motor cannibalised from the windscreen wiper of a Lancaster bomber. Bits of filed-down Perspex and fishing line were used to channel the coloured dyes that mimicked the flow of income round the economy into consumer spending, taxes, investment and exports. A.W. 'Bill' Phillips and Walter Newlyn, who helped piece the machine together at the London School of Economics towards the end of the 1940s, experimented with treacle and methylated spirits before deciding that coloured water was the best way of displaying the way money circulates around the economy.

Water flows through a series of clear pipes, mimicking the way that money flows through the economy. The machine lets you see (literally) what happens if you lower tax rates or increase the money supply or whatever; just open a valve here or pull a lever there and the machine sloshes away, showing in real time how the water levels rise and fall in various tanks representing the growth in personal savings, tax revenue and so on. This device was state of the art in the 1950s, but it looks hilarious now, with all its plumbing and noisy pumps. One early demonstration of the machine displayed the difficulties that can arise when monetary and fiscal policy are not synchronised. Phillips asked one of his students to be chancellor of the exchequer and control taxes and spending; another to be governor of the Bank of England and control interest rates. Predictably, the policies were uncoordinated, and the upshot was that water overflowed on to the floor.



James Meade with Phillips' and Newlyn's device for demonstrating the circular flow of income

Read more about the Phillips machine in Tim Harford's book, *The Undercover Economist Strikes Back*, which is also available on a BBC radio podcast.

SECTION 10.1 SUMMARY

- National income is the flow of new output produced in an economy in a particular time period.
- Changes in real national income must not be confused with changes in nominal national income.
- A circular flow model of the economy shows the flows of income and spending around the economy.
- The circular flow model can illustrate injections into and leakages (or withdrawals) from income circulating round the economy and equilibrium national income.
- Economists use the letters *Y*, *C*, *S*, *I*, *G*, *T*, *X* and *M* as shorthand, respectively, for the macroeconomic variables income, consumption, saving, investment, government spending, taxation, exports and imports.
- In a simplified circular flow model of the economy, equilibrium national income occurs when $S = I$.
- In a more realistic circular flow model of the economy containing a government sector and an overseas sector, equilibrium national income occurs when $S + T + M = I + G + X$.

KEY TERMS

aggregate demand total planned spending on real output in the economy at different price levels.

reflationary policies policies that increase aggregate demand with the intention of increasing real output and employment.

aggregate supply the level of real national output that producers are prepared to supply at different average price levels.

10.2 Aggregate demand and aggregate supply analysis

In the last 40 years or so, the *AD/AS* model has become the preferred theoretical framework that many economists use for explaining macroeconomic issues. For example, the model is useful for analysing the effect of an increase in **aggregate demand** (*AD*) upon the economy. This addresses a key issue: will expansionary fiscal policy and/or monetary policy increase real output and jobs (**reflationary policies**), or will the price level increase instead? As we will explain, the answer to this key macroeconomic question depends to a large extent on supply-side factors as represented by the shape of the **aggregate supply** (*AS*) curve, in both the short run and the long run.

The meaning of aggregate demand

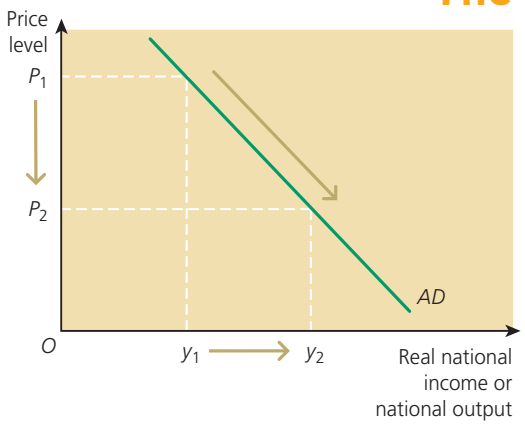


Figure 10.4 An aggregate demand (*AD*) curve

Aggregate demand (*AD*) is the total planned spending on the goods and services produced within the economy in a particular time period, such as a year. An *AD* curve is illustrated in Figure 10.4. The four sources of spending included in aggregate demand, each originating in a different sector of the economy (households, firms, the government sector and the overseas sector), are shown in the equation:

$$AD = C + I + G + (X - M)$$

where *C*, *I*, *G*, *X* and *M* are the symbols used respectively for planned consumption, investment, government spending, exports and imports.

STUDY TIP

Make sure you understand and can apply the aggregate demand equation:

$AD = C + I + G + (X - M)$

See also section 10.3 of this chapter on the determinants of aggregate demand — that is, what determines the monetary value of each component of aggregate demand.

Although it slopes downwards, this is not for the same reason as the demand curve covered in Chapter 3. One reason why the *AD* curve slopes downwards is that at lower price levels, exports will become more price competitive and are likely to rise — leading to an increased amount of spending, and increased *AD*.

As with the demand curve covered in Chapter 3, it is important that you can identify and distinguish between those factors that shift the *AD* curve leftwards or rightwards, and those that move us along the *AD* curve (usually connected with movements in the aggregate supply curve).

TEST YOURSELF 10.6

Distinguish between aggregate demand and national expenditure.

STUDY TIP

Make sure you don't confuse aggregate demand with national expenditure in your exam. Both are macroeconomic concepts, related to demand in the whole economy, or the economy in aggregate. However, aggregate demand measures *planned* spending, whereas national expenditure measures *realised* or *actual* spending, which has already taken place — for example, UK national expenditure in 2022.

QUANTITATIVE SKILLS 10.1

Worked example: components of aggregate demand

Table 10.1 shows the components of aggregate demand for an economy in 2022 and 2023.

Table 10.1 Components of aggregate demand, 2022 and 2023 (£bn)

Year	Government and private consumption expenditure	Government and private investment expenditure	Exports	Imports
2022	1,000	150	300	200
2023	1,200	170	250	220

(a) What was the change in the level of aggregate demand between 2022 and 2023?

The information in Table 10.1 relates to the aggregate demand equation:

$AD = C + I + G + (X - M)$

Adding together the information in the columns, in 2022:

AD = £1,000bn + £150bn + £300bn – £200bn = £1,250bn

In 2023:

AD = £1,200bn + £170bn + £250bn – £220bn = £1,400bn

This means that between 2022 and 2023 aggregate demand increased by £150 billion.

(b) What was the change in net export demand between 2022 and 2023?

In 2022 net export demand $(X - M)$ was £300bn – £200bn, which was £100 billion.

In 2023 net export demand $(X - M)$ was £250bn – £220bn, which was £30 billion.

There was a balance of trade surplus in both years. The trade surplus fell by £70 billion between 2022 and 2023.

The meaning of aggregate supply

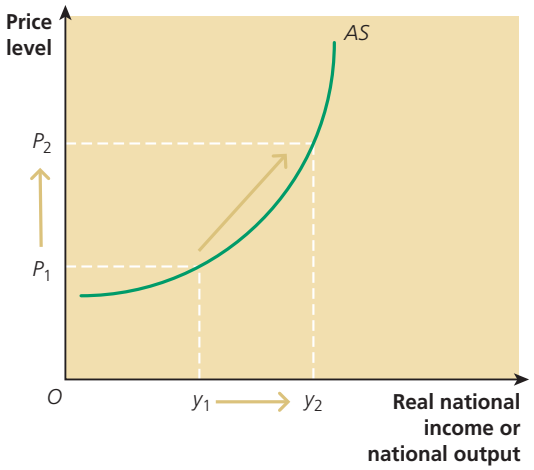


Figure 10.5 An aggregate supply (*AS*) curve

Aggregate supply (*AS*) is the level of real national output that producers are prepared to supply at different average price levels. An upward-sloping *AS* curve is illustrated in Figure 10.5.

The upward-slope of the *AS* curve is explained by two microeconomic assumptions about the nature of firms. These are:

- All firms aim to maximise profits.
- In the short run, the cost of producing extra units of output increases as firms produce more output.

At the average price level P_1 in Figure 10.5, the level of real output that all the economy's firms are willing to produce and sell is y_1. To persuade the firms it is in their interest to produce the larger output of y_2, the price level must rise. This is because higher prices are needed to create the higher sales revenues needed to offset the higher production costs that firms incur as they increase output, so that profits do not fall. In Figure 10.5, the average price level has to rise to P_2 in order to create conditions in which profit-maximising firms are willing and able to supply more output. If prices do not rise, it is not profitable to increase supply. Without a higher price level, profit-maximising firms, taken in aggregate, will not voluntarily increase the supply of real output.

TEST YOURSELF 10.7
Distinguish between aggregate demand and aggregate supply.

STUDY TIP
At A-level we do not consider the gradient of the *AD* curve to matter — as long as it is downward sloping from left to right.

Changes in the price level are shown by movements along the *AD* and *AS* curves

If you look back at Figure 10.4, you will see that as the price level in the economy falls from P_1 to P_2, the aggregate demand for real output increases. Likewise, as we have just explained, the aggregate supply curve in Figure 10.5 shows that as the price level in the economy rises from P_1 to P_2, firms are willing to supply more real output.

Factors that shift the *AD* curve and the *AS* curve

If either the *AD* curve or the *AS* curve shifts to a new position, the level of national income will change. Figure 10.6(a) shows the effect of a rightward shift of the *AD* curve, with the level of income moving from point *X* to point *Z*. By contrast, Figure 10.6(b) shows the effect of a leftward movement of the *AS* curve, with the level of income moving from point *X* to point *V*.

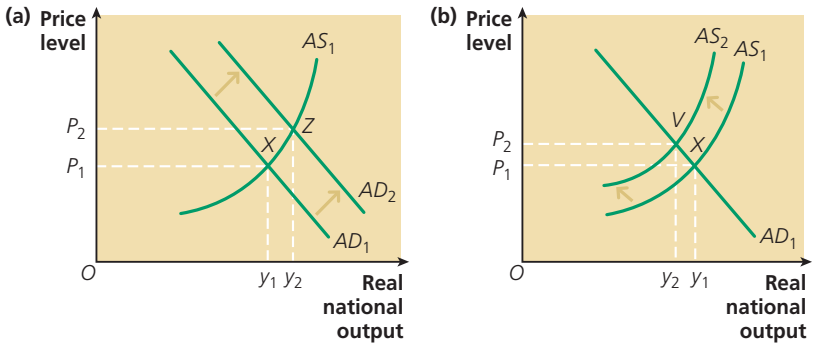


Figure 10.6 Shifts of, and adjustments along, *AD* and *AS* curves, showing movements to new levels of income

An *AD* curve will shift to a new position if there is a change in the value of any of the components of aggregate demand which we described earlier. Remember, these are consumption, investment, government spending, exports and imports. For example, a rise in business confidence could cause investment (by businesses) to increase, or a fall in interest rates could cause both consumption and investment to increase — which would shift the *AD* curve rightwards.

The *AS* curve is constructed under the assumption that all the determinants of aggregate supply *other than the price level* remain unchanged. If any of these determinants change, the *AS* curve shifts to a new position. The curve can shift either to the right (an increase in aggregate supply) or to the left (a decrease in aggregate supply). Figure 10.6(b) illustrates a decrease in aggregate supply, which results in inflation, with the price level increasing from P_1 to P_2. Increases in *SRAS* are caused by decreases in the costs of production, such as a fall in wage rates or an increase in productivity.

TEST YOURSELF 10.8
Distinguish between an adjustment along, and a shift of, an aggregate supply curve.

SYNOPTIC LINK
Don't confuse *AD* and *AS* with the microeconomic *D* and *S* concepts. The curves for *AD* and *(SR)AS* may look very similar to the *D* and *S* curves covered in Chapter 3. However, the determinants of *AD* and *AS* are quite different from the determinants of the *D* and *S* concepts.

STUDY TIP
An increase or decrease in *AD* is shown as a shift in the *AD* curve. This will lead to a movement along the *AS* curve — either an expansion of *AS* (if *AD* increases) or a contraction of *AS* (if *AD* decreases).

An increase or decrease in *AS* is shown as shift in the *AS* curve. This also leads to a movement along the *AD* curve — an expansion of *AD* (if *AS* increases) or a contraction of *AD* (if *AS* decreases).

As with microeconomic demand and supply curves, don't confuse shifts in curves with movements along the curves.

As a practice exercise, try drawing diagrams to illustrate the events just described.

Introducing the long-run aggregate supply (*LRAS*) curve

KEY TERM
long-run aggregate supply the real output that can be supplied when the economy is on its production possibility frontier. This is when all the available factors of production are employed and producing at their 'normal capacity' level of output.

Although we have not named them as such, the aggregate supply curves considered so far in this chapter are short-run aggregate supply (*SRAS*) curves. However, we now introduce the economy's **long-run aggregate supply** (*LRAS*) curve.

We have seen that in the short run, the aggregate supply or real output depends on the average price level in the economy. Other things remaining constant, firms are only prepared to supply more output if the price level rises. However, in the long run, aggregate supply is not influenced by the price level. Long-run supply reflects the economy's production potential. It is the maximum level of output that the economy can produce when the economy is on its production possibility frontier.

TEST YOURSELF 10.9
Why do *SRAS* and *LRAS* curves have different slopes?

Producing on the production possibility frontier means that there is no spare capacity or idle resources in the economy. Thus, the vertical *LRAS* curve in Figure 10.7 is located immediately above the maximum level of physical output that the economy can produce. As Figure 10.7 shows, once the economy produces at full capacity, an increase in aggregate demand from AD_1 to AD_2 increases the price level from P_1 to P_2, but real output remains unchanged.

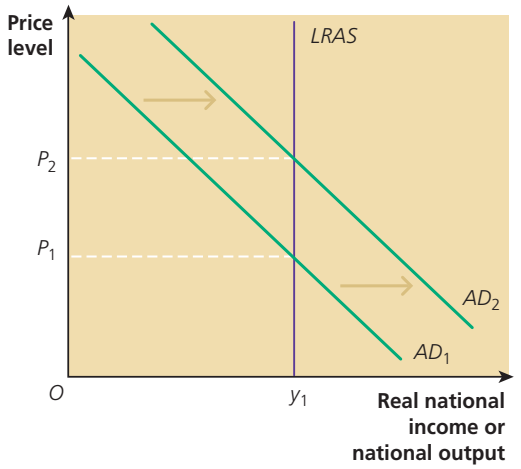


Figure 10.7 A shift of aggregate demand along a vertical *LRAS* curve

Underlying economic growth shown by a rightward shift in the *LRAS* curve

Short-run growth, which we will examine in Chapter 11 in the context of economic cycles, results mostly from temporary fluctuations in aggregate demand. By contrast, the underlying rate of growth, or trend growth rate, is the long-run average growth rate for a country over a period of time. Underlying growth is determined by the same factors that determine the position of the economy's production possibility frontier: the quantities of capital and labour and other factors of production in the economy, and technical progress. These also determine the position of the *LRAS* curve.

An increase in the quantity of available factors of production including labour, and improvements in technology that increase the productivity of labour, capital or land, shift the economy's production possibility frontier outwards. For the same reasons, the economy's *LRAS* curve shifts right. The shifts of the production possibility frontier and the *LRAS* curve to the right are both illustrated in Figure 10.8.

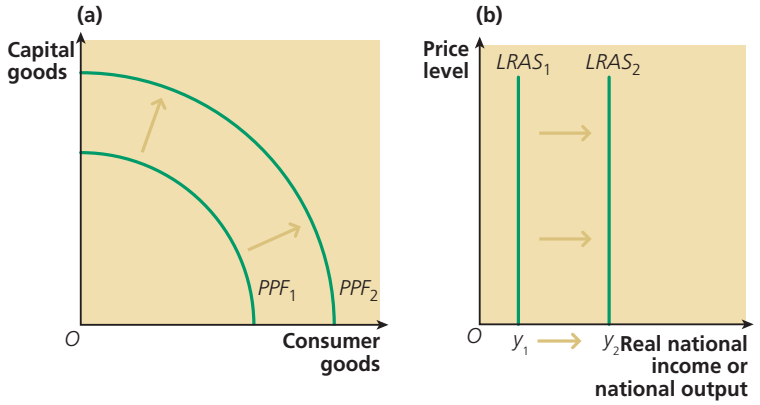


Figure 10.8 Underlying growth shown by an outward movement of an economy's production possibility frontier and by a rightward movement of its *LRAS* curve

STUDY TIP

You should appreciate that long-run economic growth can be shown either by an outward movement of the economy's production possibility frontier or by a rightward shift of the *LRAS* curve. Short-run economic growth comes from an increase in *AD* (or, occasionally, an increase in *SRAS*).

A second look at the concept of equilibrium national income

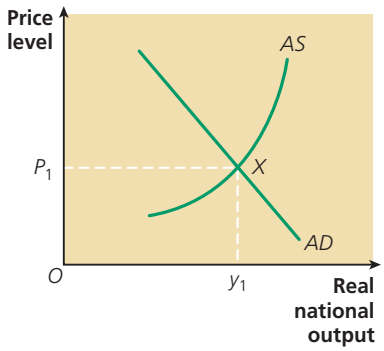


Figure 10.9 Equilibrium national income occurring when *AD* = *AS*

There are two ways of approaching the concept of equilibrium national income (or macroeconomic equilibrium). We have already seen how national income is in equilibrium when planned injections into the circular flow of income equal planned withdrawals of demand from the flow of income circulating round the economy.

Ignoring for the time being the distinction between short-run and long-run aggregate supply, the second approach defines equilibrium national income in the *AD/AS* model of the economy. Equilibrium national income occurs when the aggregate demand for real output equals the aggregate supply of real output, i.e. where *AD* = *AS*. This is illustrated in Figure 10.9 at point *X*, where the *AD* curve intersects the *AS* curve. The equilibrium level of real output is y_1, and the equilibrium price level is P_1.

Demand-side and supply-side economic shocks

KEY TERM

economic shock an unexpected event hitting the economy. Economic shocks can be demand-side or supply-side shocks (and sometimes both), and unfavourable or favourable.

An **economic shock** is a sudden unexpected event hitting the economy, disturbing either aggregate demand (a demand shock) or aggregate supply (a supply shock). In some cases, an outside shock may affect both aggregate demand and aggregate supply. Thus, looking at the UK economy, the outbreak of a war in the Middle East, for example, may affect demand by causing a sudden collapse in consumer and business confidence, and aggregate supply via its effect on the supply and price of crude oil.

Since 2007, there have been four major economic shocks that have affected the UK economy. Some of these have affected the demand side of the economy, some have affected the supply side of the economy and some have affected both. These are the financial crisis of 2007–08, the UK leaving the European Union (the shock arguably occurred once the UK fully left the EU in 2021), the Covid-19 pandemic of 2020 (and beyond), and the war in Ukraine beginning in 2022.

STUDY TIP

When deciding which of the *AD* and *AS* curves will shift, consider the following:

- *AD* is determined by factors affecting national spending.
- *SRAS* is determined by factors that affect costs of production and the profitability of production.
- *LRAS* is determined by the economy's productive capacity.

APPLICATION OF ECONOMICS IN THE REAL WORLD 10.1

Economic shock caused by the El Niño effect

Above-average ocean surface temperatures develop every three to seven years off the Pacific coast of South America and last about two years, causing major climatological changes around the world. This is the El Niño effect. The 2015–16 El Niño was one of the most severe events in the past 50 years and the largest since the 1997–98 El Niño that shocked global food, water, health, energy and disaster-response systems. There are also less frequent cold events, called La Niña, when temperatures in the winter are higher than normal in the south, but cooler than normal in the north.

Economists are increasingly interested in the relationship between climate — temperature, precipitation, storms and other aspects of the weather — and economic performance. The extreme weather conditions associated with El Niño can constrain the supply of rain-driven agricultural commodities, lead to higher food prices and inflation, and may trigger social unrest in commodity-dependent countries that rely primarily on imported food.

An El Niño typically brings drought to the western Pacific (including Australia), rains to the equatorial coast of South America, and storms and hurricanes to the central Pacific. These changes in weather patterns have significant effects on agriculture, fishing and construction industries, as well as on national and global commodity prices.

Australia, Chile, India, Indonesia, Japan, New Zealand and South Africa face a short-lived fall in economic activity in response to a typical El Niño shock. In other countries, the effect on the climate actually boosts GDP. In the USA, for example, it is linked with a fall in hurricanes and improved outputs of certain crops. Many countries experience short-term inflation pressures following an El Niño shock (its magnitude increasing with the share of food in the consumer prices index (CPI) basket), while the prices of energy and non-fuel commodities (goods which are, in effect, identical, so are interchangeable) also rise around the world.

Climate change and global warming in sub-Saharan economies have led to severe drought and desertification, which is dramatically increasing the area of the Sahara desert. A number of countries in Africa are experiencing economic crisis and increasing poverty due to civil war, drought and famine. This means many people choose to emigrate in search of better economic prospects.

Source: IMF, March 2016

Follow-up questions

1. Would you classify the El Niño effect as a demand-side or as a supply-side economic shock, or as a mix of the two? Justify your conclusion.
2. Explain how the El Niño effect has affected global inflation.
3. In what ways, if any, would the 2018 and 2022 hot summers in the UK have beneficial or adverse effects on the UK economy?
4. Do you agree that climate change and global warming in sub-Saharan Africa are having much more significant adverse effects on world economic performance than the El Niño effect in the Pacific Ocean? Justify your answer.

STUDY TIP

You should build up knowledge of four or five economic shocks which have affected the UK and global economies in recent years. Consider whether these are demand-side or supply-side shocks (or both).

SECTION 10.2 SUMMARY

- The *AD/AS* model can be used as well as the circular flow model to show equilibrium national income (or macroeconomic equilibrium) in the economy.
- The *AD* curve shows total planned spending on real national output at different price levels.
- The *SRAS* curve shows how much output producers are prepared to supply at different price levels.
- It is important to distinguish between a *shift* of an *AD* or *AS* curve and the resulting *movement along* the curve that does not shift.
- A shift of the *AD* curve to the right is known as an increase in aggregate demand, while a shift to the left is a decrease in aggregate demand.
- A shift of the *SRAS* curve to the right is known as an increase in aggregate supply, while a shift to the left is a decrease in aggregate supply.
- Long-run aggregate supply (*LRAS*) is the maximum level of output that the economy can produce when on its production possibility frontier.

10.3 The determinants of aggregate demand

The components of aggregate demand

As mentioned earlier in the chapter, consumption, investment, government spending and net export demand (spending on UK exports by residents of other countries minus spending on imports by UK residents) form the components of aggregate demand.

We will now look in detail at the first two of the components of aggregate demand: consumption and investment.

SYNOPTIC LINK

We will explain government spending's role as a component of aggregate demand in Chapter 13, section 13.1, on fiscal policy. The role of exports and imports will be explained in Chapter 14, section 14.3, concerning the balance of payments on current account.

The determinants of consumption

Aggregate consumption is spending by all the households in the economy on consumer goods and services. Whenever members of households make decisions about whether or not to spend on consumer goods, they are simultaneously deciding whether or not to save. A determinant of consumption is therefore also a determinant of household saving. If we assume a closed economy — assuming that there are no exports or imports — and that there is no taxation, then at any level of income, households can only do two things with their income: spend it or not spend it. Spending income is consumption, whereas not spending income is saving.

The determinants of consumption are described below.

TEST YOURSELF 10.10

What is the difference between the macroeconomic concept of consumption and the microeconomic concept of market demand?

KEY TERM
rate of interest the reward for lending savings to somebody else (e.g. a bank) and the cost of borrowing.

Interest rates

The **rate of interest** rewards savers for sacrificing current consumption, and the higher the rate of interest, the greater the reward. Thus, at any particular level of income, the amount saved will increase as the real rate of interest (the rate of interest adjusted for the inflation rate) rises and the amount consumed will fall.

QUANTITATIVE SKILLS 10.2

Worked example: calculating the real rate of interest

If the rate of interest is currently 4.5% and the inflation rate is currently 6.2%, what would be the real rate of interest?

real rate of interest = nominal rate of interest – current inflation rate
= 4.5% – 6.2% = –1.7%

Level of income

An important, and perhaps the main, determinant of both consumption and saving is the level of income. This is often called the Keynesian theory of consumption and saving. In 1936, in his explanation of the causes of the Great Depression, John Maynard Keynes wrote:

> The fundamental psychological law, upon which we are entitled to depend with great confidence … is that men are disposed, as a rule and on average, to increase their consumption as their income increases, but not by as much as the increase in their income.

In Keynes's view, as income rises, although absolute consumption rises, consumption falls as a fraction of total income, while the fraction saved increases. Therein lies the cause of recessions, according to Keynes: too much saving and too little spending.

Attempts to save more by households could lead to overall income falling — as less spending by a household will mean less income is earned by others. The multiplier effect (see page 353) will mean that overall income will fall. This is sometimes referred to as the paradox of thrift — where an attempt to save more in the economy actually leads to lower income and lower overall savings.

Expected future income

The Keynesian consumption theory just explained is sometimes also called the 'absolute income' consumption theory because it assumes that the most important influence on consumption is the *current* level of income. However, the current level of income in a particular year may have less influence on a person's planned consumption than some notion of *expected* income over a much longer time period, perhaps extending over the individual's remaining lifetime or life cycle.

People plan a large part of their savings on the basis of a long-term view of their expected lifetime or permanent income, and of likely spending plans over the remaining length of an expected life cycle — this is the **life-cycle theory of consumption**. Temporary fluctuations in yearly income generally have little effect on forms of saving such as contributions to pension schemes and the purchase of life insurance policies.

KEY TERM
life-cycle theory of consumption a theory that explains consumption and saving in terms of how people expect their incomes to change over the whole of their life cycles.

TEST YOURSELF 10.11
Distinguish between current income and life-cycle income.

TEST YOURSELF 10.12
What is the difference between a stock and a flow?

Wealth

The stock of personal wealth, as well as the flow of income, influences consumption and saving decisions. In countries such as the UK and the USA, houses and shares are the two main forms of wealth asset that people own. An increase in house prices usually causes homeowners to consume more and to save less from their current flow of income, partly because the wealth increase 'does their saving for them'. Rising house prices generally increase the amount of borrowing taking place in an economy, partly through house buyers taking out large mortgages to secure a house purchase. The additional borrowing finances extra consumption, not only on the houses themselves but on items such as furniture and new bathrooms and kitchens.

Additionally, some owner-occupiers who do not wish to move house may take out a larger mortgage on the house they live in, and then spend what they borrow: for example, on new cars and holidays. This is called equity release. In this context, the word 'equity' means 'wealth'. Borrowing against the value of a house reduces the amount of equity 'locked up' in people's houses.

STUDY TIP
Equity has a number of different meanings in economics. Here it refers to the wealth of an individual. However, it can also refer to the capital of a business owner (covered in Chapter 12) and to the idea of fairness (covered in Chapter 7).

Rising house prices induce a 'feel-good' factor among property owners, which leads to a consumer spending spree in the shops. Conversely, falling house prices have the opposite effect, increasing uncertainty and precautionary saving via a 'feel-bad' factor.

When share prices rise, share owners also become wealthier and may finance extra consumption by using borrowed funds. However, this effect is less noticeable in the UK, where houses rather than shares are the main household wealth asset. Stock market crashes have the opposite effect to booming share prices, reducing the wealth of shareholders, and consequently their consumption.

In the summer of 2007, a dramatic fall in share prices occurred in the USA, and then in other countries, including the UK. A year later, in 2008, falling house and share prices, accompanied by a collapse in consumer and business confidence, ushered in recession in the USA and then in other countries such as the UK. The collapse in share prices in early 2020 did not have the same effect, as this was the result of an anticipated slowdown due to the Covid-19 pandemic. Share prices largely recovered in value over the next 18 months.

Consumer confidence

The state of consumer confidence is closely linked to people's views on expected income and to changes in personal wealth. When consumer optimism increases, households generally spend more and save less, whereas a fall in optimism (or a growth in pessimism) has the opposite effect.

Governments try to boost consumer (and business) optimism to ward off the fear of a collapse in confidence by 'talking the economy up' and by trying to enhance the credibility of government economic policy. If the government is

optimistic about the future, and people believe there are good grounds for this optimism, then the general public will also be optimistic and confident about the future. However, if people believe the government is pursuing the wrong policies, or if an adverse economic shock hits the economy in a way that the government can't control, confidence can quickly dissipate.

The availability of credit

KEY TERMS

availability of credit funds available for households and firms to borrow.

credit crunch occurs when there is a lack of funds available in the credit market, making it difficult for borrowers to obtain finance and leading to a rise in the cost of borrowing.

Besides the rate of interest, other aspects of monetary policy, such as controls on bank lending, affect consumption. If credit is available easily and cheaply (i.e. where interest rates are low and it is fairly easy for people and businesses to borrow money), consumption increases as people supplement current income by borrowing on credit created by the banking system. Conversely, a tight monetary policy reduces consumption.

The financial crisis that occurred in 2007 and 2008 had this effect in the UK and the USA. It arose from bad debts in the US sub-prime market — the parts of the financial markets that deal with lending to those with poor credit scores (i.e. those who are a bigger risk to lend money to and are more likely to be unable to repay their debts). In the so-called **credit crunch**, interest rates rose and the supply of credit dried up, with banks refusing to supply applicants with new credit cards or mortgages. Banks were increasingly worried about running short of money for themselves and therefore were more reluctant to lend money out.

Distribution of income

KEY TERM

distribution of income the spread of different incomes among individuals and different income groups in the economy.

Consumption and saving are also influenced by the **distribution of income** within an economy. Rich people save a greater proportion of their income than poor people; redistribution of income from rich to poor therefore increases consumption and reduces saving.

Expectations of future inflation

It is not easy to predict the impact of inflation on consumption. However, uncertainty caused by fears of rising inflation increases precautionary saving and reduces consumption. It may also, however, have the opposite effect. Households may decide to bring forward consumption decisions by spending now on consumer durables such as cars or television sets, thereby avoiding expected future price increases. People may also decide to borrow to finance the purchase of houses if they expect property prices to appreciate at a rate faster than general inflation. In this situation, and particularly if the real rate of interest is low or negative, people often decide to buy land, property and other physical assets such as art and antiques as a 'hedge' against inflation (because they keep their value), in preference to saving through the purchase of financial assets.

STUDY TIP

Make sure you appreciate how consumer confidence and business confidence affect economic activity, and consider whether they affect *AD* or *AS* (and through which component of *AD* or *AS*).

The determinants of savings

Saving is a decision by people to postpone their consumption. Whenever members of households make decisions whether or not to spend on consumer goods, they are simultaneously deciding whether or not to save. Therefore all the factors we have listed in the previous section as determinants of consumption are also determinants of saving. Consider the following equation:

$$Y = C + S$$

The equation is telling us that, if we ignore taxation and spending on imports, people can only do two things with their income: spend it on consumption or save. We can rewrite the equation as:

$$S = Y - C$$

This is the definition of saving, namely income which is not spent on consumption.

TEST YOURSELF 10.13
Explain the difference between saving and consumption.

EXTENSION MATERIAL

The personal savings ratio and the household savings ratio

The personal savings ratio measures the *actual* or *realised* saving of the personal sector as a ratio of total personal sector disposal income:

$$\text{personal savings ratio} = \frac{\text{realised or actual personal saving}}{\text{personal disposable income}}$$

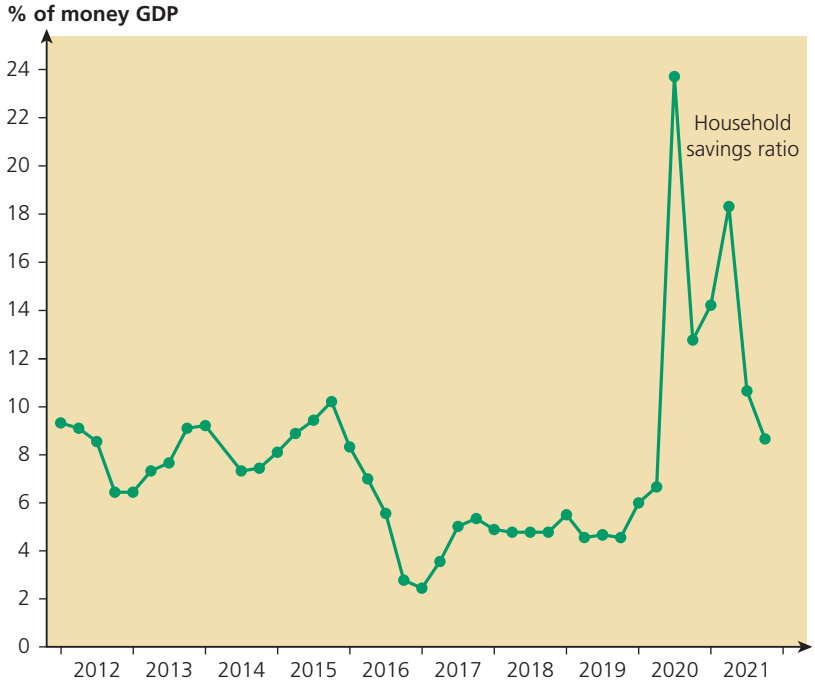
The household savings ratio is used in a similar way. It measures households' realised saving as a ratio of their disposable income (including benefit payments) after deductions for tax.

Economists and the government are interested in how much of their incomes people plan to save and to consume in the near future, as this provides important information about what lies ahead for the state of aggregate demand. It is difficult to measure people's plans accurately, so the personal savings ratio calculated for the most recent past period is generally used as an indicator of what people wish to do in the future.

The savings ratio had moved between 7 and 10% for the period 2012 to 2016 (see Figure 10.10). However, from 2016 onwards it fell below 5% and remained around that level until 2020. The Covid-19 pandemic beginning in early 2020 was associated with a rise in the savings ratio. During the pandemic's main impact in 2020 and 2021, the savings ratio reached very high levels (peaking at over 20%). This was probably the result of a number of factors, such as increased uncertainty and insecurity over people's future jobs and income prospects. A more obvious factor causing the rise in savings was the restrictions placed on people's shopping habits. Spending on holidays, and visits to restaurants and pubs were clearly curtailed during the early part of the pandemic, which meant people saved much more of their surplus income than they normally would.

One of the factors affecting household savings is people's uncertainty about the future. People who fear job losses, or who expect to suffer a fall in future income, are likely to save more for precautionary reasons. They want a 'nest egg' (a sum of money to be used in the future, often for a particular purpose) to protect themselves from the loss of future income.

This means that savings ratios generally rise in recessions, unless overridden by other factors, such as the need to borrow more so as to maintain spending levels and current standards of living and lifestyles.



Source: ONS

Figure 10.10 Changes in the UK household savings ratio, 2012–21

APPLICATION OF ECONOMICS IN THE REAL WORLD 10.2

The 'credit crunch' and the financial crisis

The financial crisis which hit the world economy in 2007 started in the US housing market where banks had lent money to **sub-prime borrowers**. When these borrowers couldn't repay their loans, or even the interest on them, the banks were left with bad debts, as were other financial institutions that had bought 'packages' of bad debt from the banks.



Customers queue to withdraw their savings from Northern Rock

This infected the entire financial system and meant that banks stopped lending to each other — creating the **credit crunch**. The first UK bank to be affected by the seizing up of financial markets was Northern Rock, which had to be rescued by the UK government. (A few years later, the government sold the bank to Virgin Money.) A lack of mortgages meant the market began to stagnate and the properties that did change hands went for less than they would have done a few months previously. This increased the amount of **negative equity** in the economy.

Follow-up questions

1. Explain the bold terms: sub-prime borrowers; credit crunch; negative equity.
2. Find out how the UK housing market has fared since 2019.
3. Access Deloitte's chief economist Ian Stewart's article 'How to spot a bubble' at **https://blogs.deloitte.co.uk/mondaybriefing/2019/03/how-to-spot-a-bubble.html**. Summarise the main message of the article.
4. It is generally agreed that the 2007 financial crisis was a major cause of the recession which hit the UK economy a year later in 2008. Why should a financial crisis be a cause of recession?

The difference between saving and investment

Economists make a clear separation between saving and investment, even though in everyday language the two terms are often used interchangeably. Whereas saving is simply income that is not spent on consumption, investment is spending by firms on capital goods such as machines and office equipment. *Physical investment* in new capital goods should not be confused with *financial investment*. The latter is the demand for financial assets such as shares and bonds, which is a form of saving.

As a simplification, economists often assume that households make saving decisions, while firms make investment decisions. Firms invest when they buy capital goods such as machinery. However, firms also save: for example, when they store profits in a bank account without spending them.

STUDY TIP

Make sure you don't confuse investment with saving. In everyday usage, investment can seem to mean the same as savings. In an exam, investment and savings are quite different. Investment is spending by businesses on additions to capital stock or to inventory levels, whereas savings represents household disposable income not spent on consumption.

The determinants of investment

A country's gross investment includes two parts: replacement investment (to make good depreciation or capital consumption), which simply maintains the size of the existing capital stock by replacing worn-out capital; and net investment, which adds to the capital stock, thereby increasing productive potential. Along with technical progress, net investment is one of the engines of economic growth.

The new capital goods created by new fixed investment often have an economic life extending many years into the future. The future is always uncertain and the further we go into the future, the greater the uncertainty. When deciding whether or not to go ahead with a fixed investment project in, for example, new machinery, firms need to form views on expected future sales revenue attributable to the investment project and expected future costs of production, resulting both from the rate of interest paid for borrowing the funds to finance the initial investment, and from the future maintenance costs.

Apart from the rate of interest, other factors that influence investment decisions are described below.

The relative prices of capital and labour

When the price of capital rises (e.g. when the prices of capital goods or rates of interest rise), in the long run firms adopt more labour-intensive methods of production, substituting labour for capital. A decrease in the relative prices of capital goods has the opposite effect. If the price of capital goods or interest rates falls, firms switch to capital-intensive methods of production, so investment increases.

The nature of technical progress

Technical progress can make machinery obsolete or out of date. When this happens, a machine's business life becomes shorter than its technical life: that is, the number of years before the machine wears out. A sudden burst of technical progress may cause firms to replace capital goods early, long before the end of the equipment's technical life.

The adequacy of financial institutions in supplying investment funds

As mentioned, many investments in fixed capital goods are long-term investments that yield most of their expected income several years into the future. These investments may be difficult to finance because of the inadequacy of the financial institutions that provide investment funds. Banks have been criticised for favouring short-term investments and being reluctant to provide the finance for long-term investments. Likewise, the stock market may favour short-termism over long-termism, although in recent years the growth of new methods of lending to firms has provided an important source of medium- to longer-term finance.

Government support

Governments also lend funds to firms to finance investment projects, although at the same time they tax firms: for example, by levying corporation tax (a UK tax on company profits). Arguably, however, when choosing whether to invest in or support investment projects, governments may be better at 'picking losers' than 'picking winners'. In the past, UK governments have sometimes provided investment funds to rescue jobs in loss-making and uncompetitive industries that perhaps should have been allowed to continue their decline. Government ministers and their civil servants may make bad investment decisions because, unlike entrepreneurs, they do not face the risk of being bankrupted as a result of poor decision making.

TEST YOURSELF 10.14
Banks function as financial intermediaries. Explain what this means.

STUDY TIP
It is important to remember that capital is a *stock* concept, but investment is a *flow*: in other words, the flow of investment expenditure over time adds to the size of the total capital stock at a point in time.

The accelerator process

KEY TERM
accelerator a change in the level of investment in new capital goods induced by a change in national income or output. The size of the accelerator depends on the economy's capital–output ratio.

The **accelerator** process stems from a simple and mechanical assumption that firms wish to keep a relatively fixed ratio, known as the capital–output ratio, between the output they are currently producing and their existing stock of fixed capital assets. If output grows by a constant amount each year, firms invest in exactly the same amount of new capital each year to enlarge their capital stock so as to maintain the desired capital–output ratio. From year to year, the level of investment is therefore constant. When the rate of growth of output *accelerates*, investment increases as firms take action to enlarge the stock of capital to a level sufficient to maintain the desired capital–output ratio. When the rate of growth of output *decelerates*, investment declines.

EXTENSION MATERIAL

A numerical example of the accelerator

To illustrate the accelerator principle, we will assume that the economy's capital–output ratio is 4:1, or simply 4. This means that 4 units of capital are required to produce 1 unit of output. We will also assume that the level of *current* net investment in fixed capital depends on the change in income or output in the *previous* year:

$$I = v(\Delta Y)$$

where Δ refers to the change in a variable, in this case, Y

or:

$$I_t = v(Y_t - Y_{t-1})$$

where I_t is net investment this year, Y is current national income, Y_{t-1} is national income last year and v is the capital-output ratio. The *capital–output ratio*, v, is also known as the *accelerator coefficient*, or simply as the *accelerator*. We will also assume no replacement investment is needed, and that the average capital-output ratio in the economy stays at 4.

Given these assumptions, consider the numerical example of the accelerator principle shown in Table 10.2.

Table 10.2 The accelerator: numerical example

Year	Net investment (£bn)		Capital output ratio (£bn)		Current income this year (£bn)		Income last year (£bn)
2020	40	=	4	×	(100	–	90)
2021	40	=	4	×	(110	–	100)
2022	80	=	4	×	(130	–	110)
2023	40	=	4	×	(140	–	130)

In each of the years from 2020 to 2023 national income grows. Between 2019 (the year preceding the data in the table) and 2020, income grows by £10 billion. Via the capital-output ratio, the £10 billion income growth induces net investment of £40 billion in 2020. The size of the capital stock increases by £40 billion, which enables the desired capital–output ratio to be maintained at the now higher level of income. In row 2, income continues to grow by £10 billion, so investment in 2021 remains at £40 billion.

However, the next year is different. In 2022, shown in row 3, the growth of income speeds up or *accelerates*, doubling from £10 billion to £20 billion. Investment also doubles from £40 billion to £80 billion to maintain the value of the capital–output ratio. Thus, a £10 billion increase in income induces a £40 billion increase in investment. But in row 4, the growth of income falls back to £10 billion in 2023. Although income is still growing, net investment has fallen back to £40 billion.

This example shows how the accelerator gets its name. The figures show the rate of growth of income and output determining whether investment grows, falls or remains at a constant level.

In summary:

- If income grows by the same amount each year, net investment is constant.
- If income growth speeds up or *accelerates*, net investment increases.
- If income growth slows down or *decelerates*, net investment declines.

Thus, as firms adjust the stock of capital to the desired level, relatively slight changes in the rate of growth of income or output cause quite large absolute rises and falls in investment. The acceleration principle therefore provides an explanation of why investment in capital goods is a more volatile or unstable component of aggregate demand than consumption.

TEST YOURSELF 10.15

Households 'save', but firms 'invest'. When do firms save?

SECTION 10.3 SUMMARY

- Aggregate demand is total *planned* spending on real output of all the economic agents in the economy. As an equation, $AD = C + I + G + (X - M)$.
- Consumption by households and investment by firms are two components of aggregate demand.
- Saving, or income which is not consumed, must not be confused with investment.
- Aggregate consumption and saving are determined by the rate of interest, the levels of current and future income, wealth, consumer confidence and the availability of credit.
- Investment is determined by factors such as the rate of interest, expected returns on capital, technical progress, the availability of finance and the accelerator.
- The accelerator measures how a change in the level of investment in new capital goods is induced by a change in national income or output.
- The size of the accelerator depends on the economy's capital-output ratio.

10.4 Aggregate demand and the level of economic activity

What is economic activity?

Economic activity can mean many things, but we will think of it as centring on the production and consumption of goods and services in the economy, together with the employment of the labour, capital and other inputs that produce output.



The buying and selling of goods and services is one aspect of economic activity

Aggregate demand and economic activity

AD/AS diagrams can be used to illustrate how changes in aggregate demand affect the level of real output, but they do not directly show the employment levels at each level of real output. However, the main link between aggregate demand and employment is simple. When real output increases, firms generally have to employ more workers to produce the additional goods and services that the output increase involves. Conversely, when real output falls, less labour is required to produce the smaller amount of goods and services now being produced.

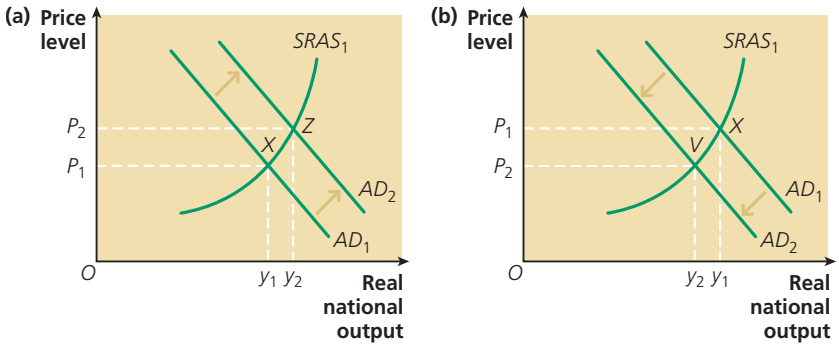


Figure 10.11 The effect of a change in aggregate demand on real output

Figure 10.11(a), which illustrates the effect of an increase in aggregate demand on real output, is the same as Figure 10.6(a). It shows how a rightward shift of the *AD* curve leads to an increase in real output. Figure 10.11(b), by contrast, shows how a leftward shift of the *AD* curve leads to a fall in real output. To put it another way, Figure 10.11(a) shows the expansionary effect of an increase in aggregate demand, while Figure 10.11(b) shows the contractionary effect of a decrease in aggregate demand. The expansionary effect is likely to increase employment; the contractionary effect is likely to lead to a fall in employment.

Two points to note about both diagrams:

- The greater the shift in aggregate demand, the greater the change in real output.
- The extent to which real output changes also depends on the steepness of the *AS* curve.

Aggregate demand and the national income multiplier

KEY TERM
multiplier the relationship between a change in aggregate demand and the resulting generally larger change in national income.

The national income **multiplier** measures the relationship between an initial change in a component of aggregate demand, such as government spending or private-sector investment, and the resulting generally larger change in the level of national income.

Suppose, for example, that government spending increases by £10 billion, but tax revenue remains unchanged. The resulting budget deficit initially injects £10 billion of new spending into the circular flow of income. This spending increases people's incomes. If we assume that everybody in the economy saves a small fraction of any income increase and spends the rest, the £10 billion

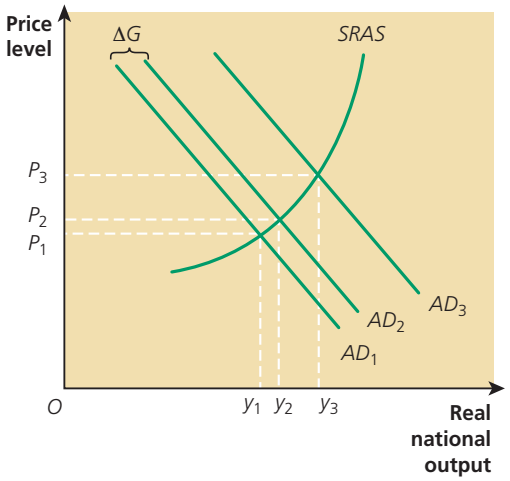


Figure 10.12 The national income multiplier illustrated on an *AD/AS* diagram

generates multiple and successively smaller further increases in income, until the next stage is so small that it can be ignored. Adding up the successive stages of income generation, the total increase in income is a multiple of the initial spending increase of £10 billion — hence the name 'multiplier theory'. If the size of the multiplier is 2.5, an increase in a component of aggregate demand, such as government spending, of £10 billion causes national income eventually to increase by £25 billion. Figure 10.12 illustrates the multiplier in an *AD/AS* diagram.

In Figure 10.12, an initial increase in government spending (ΔG) shifts the *AD* curve from AD_1 to AD_2. This then triggers the multiplier process, which leads to a further increase in aggregate demand to AD_3. If the size of the government spending multiplier is 2.5, then, as we have mentioned, the eventual increase in aggregate demand is two and a half times the size of the initial increase in government spending.

The multiplier relationship is shown in the following equation:

$$\text{multiplier} = \frac{\text{change in national income}}{\text{initial change in government spending}}$$

or:

$$\text{multiplier} = \frac{\Delta Y}{\Delta G}$$

As Quantitative Skills 10.3 shows, the size of the multiplier can be worked out, provided we possess knowledge of both the change in aggregate demand and the resulting change in national income.

STUDY TIP

Make sure you don't confuse the multiplier with the accelerator. They are easy to mix up — both concepts look at the relationship between *AD* (especially the investment component, in the case of the accelerator) and national income. The confusion can come from misunderstanding which variable changed to cause a change in the other variable.

The marginal propensities to consume and save

Before we explain the formula used for calculating the size of the multiplier, we will first introduce the marginal propensities to consume and save. The **marginal propensity to consume** (*MPC*) is the fraction of any increase in income which people plan to spend on the consumption of domestically produced goods. For example, if people, on average, plan to spend 90p of an income increase of £1 on consumption, the *MPC* is 0.9. Likewise, **marginal propensity to save** (*MPS*) is the fraction of any increase in income which people plan to save rather than spend. If the *MPC* is 0.9, the *MPS* must be 0.1, and so on.

KEY TERMS

marginal propensity to consume the fraction of any increase in income which people plan to spend on the consumption of domestically produced goods and services.

marginal propensity to save the fraction of any increase in income which people plan to save rather than spend.

QUANTITATIVE SKILLS 10.3

Worked example: calculating the size of the multiplier

In an economy, nominal national income is £2,000 billion in 2021. Government spending increases by £10 billion. The change in government spending causes nominal national income to increase to £2,050 billion in 2022.

What is the size of the government spending multiplier in relation to the change in *nominal* national income?

The increase in government spending of £10 billion causes nominal national income to increase by £50 billion between 2021 and 2022. The size of the multiplier in terms of the growth in nominal national income is thus £50 billion divided by £10 billion, which is 5. This is a quite a large multiplier.

EXTENSION MATERIAL

The different multipliers

Nested within the national income multiplier are a number of specific multipliers, each related to the particular component of aggregate demand that initially changes. Besides the government spending multiplier, there is an investment multiplier, a tax multiplier, an export multiplier and an import multiplier. Taken together, the government spending and tax multipliers are known as fiscal policy multipliers. Likewise, the export and import multipliers are foreign trade multipliers. An increase in consumption spending can also trigger a multiplier process.

The multiplier process can also work in reverse, reducing rather than increasing national income. This happens when government spending, consumption, investment or exports fall. It also happens when taxation or imports increase. This is because taxation and imports are withdrawals from the circular flow of income, rather than injections. The tax and import multipliers are always negative, meaning that an increase in taxes or imports causes national income to fall (and a decrease in taxes or imports causes national income to rise).

All the multipliers we have so far been considering are national multipliers. However, regional multipliers can also be important. A regional multiplier measures the relationship between an initial change in aggregate demand and the resulting change in regional income. Consider a situation in which central government pays a subsidy to people in Cornwall. When the recipients of the subsidy spend the extra income they receive, most of it may be spent on goods and services produced in other parts of the UK, if not on imports. The regional multiplier, which reflects extra local spending within Cornwall, is likely to be very small.

TEST YOURSELF 10.16

Which of the following is correct?

The multiplier can refer to the effect of a change in the level of:

A exports on national income

B national income on investment

C consumption on saving

D saving on imports

The multiplier formula

The formula for calculating the value of the multiplier is:

$$k = \frac{1}{1 - MPC} \text{ or } \frac{1}{MPS}$$

where k is the multiplier, *MPC* is the marginal propensity to consume, and *MPS* is the marginal propensity to save.

TEST YOURSELF 10.17
Explain the difference between the multiplier and the accelerator.

High marginal propensities to consume mean that, at each stage of the multiplier process illustrated in Figure 10.13, most new income is spent on consumption and only a small amount leaks into saving. In this situation, multipliers are relatively large. An initial increase in government spending or private investment has a larger multiplier effect than would be the case if the *MPC* were smaller. The eventual growth in income is significantly larger than the initial increase in aggregate demand. A low *MPC* (and a high *MPS*) produces the opposite effect: the eventual growth in income resulting from the multiplier process is not much larger than the initial increase in aggregate demand. In this situation, in which saving and the other leakages from the circular flow of income are quite high, multipliers are generally small, not significantly different from 1.

EXTENSION MATERIAL

The multiplier as a dynamic process

The multiplier process is essentially dynamic, taking place over time. It resembles ripples spreading over a pond after a stone has been thrown in the water. However, the ripples in a pond last only a few seconds, whereas the ripples spreading through the economy following a change in aggregate demand can last for months and even years. Figure 10.13 illustrates the ripple effect. The diagram, which shows the government spending multiplier, can easily be adapted to illustrate the investment multiplier or any other national income multiplier. The assumption of quite a small *MPC* of 0.2 is realistic for the UK and many other developed economies because a large fraction of any increase in income leaks out of the circular flow of income in taxation and spending on imports, as well as into saving. As the next paragraph explains, this reduces the size of the multiplier to the small size of 1.25.

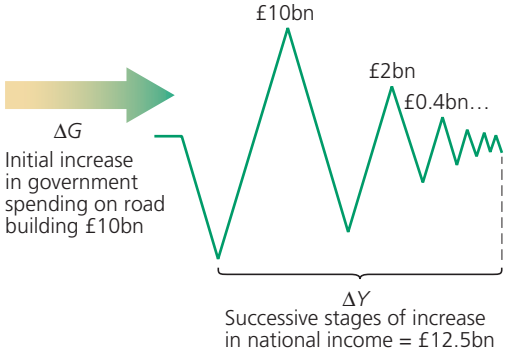


Figure 10.13 The dynamic nature of the multiplier process

In the first stage of the multiplier process illustrated in Figure 10.13, the £10 billion of extra government spending is received as income by building workers who, like everybody in the economy, spend 20p of every pound of income on consumption. (We are assuming the marginal propensity to consume (*MPC*) is 0.2 throughout the economy, which means that people plan to consume 20p out of an increase in income of £1. This means that at each stage of the multiplier process, 80% of an extra pound of income leaks into saving, taxation and imports, and is not spent on consumption.)

At the second stage of the multiplier process, £2 billion of the £10 billion income is spent on consumer goods and services, with the remaining £8 billion leaking into savings, the payment of taxes and spending on imports.

At the third stage, consumer goods sector employees spend £0.4 billion, or 0.2 of the £2 billion received at the second stage of income generation.

Further stages of income generation then occur, with each successive stage being 0.2 of the previous stage. Each stage is smaller than the preceding stage because a large part of income leaks into savings, taxation and imports at each stage of the multiplier process.

Assuming that nothing else changes in the time taken for the process to work through the economy, the eventual increase in income ΔY resulting from the initial injection of government spending is the sum of all the stages of income generation. In our example, the eventual increase in income (ΔY) is larger than ΔG, which triggered the initial growth in national income, but the multiplier is quite small (1.25) because of the size of leakages out of the circular flow of income at each stage of the multiplier process.

TEST YOURSELF 10.18

Calculate the value of the marginal propensity to save and the value of the multiplier when the value of the marginal propensity to consume is:

(a) 0.9

(b) 0.1

(c) 0.5

QUANTITATIVE SKILLS 10.4

Worked example: using the *MPC*

Calculate the size of the multiplier when, in a closed economy with no exports or imports, the size of the marginal propensity to consume (*MPC*) is (a) 0.6 and (b) 0.4.

(a) When the *MPC* is 0.6, the multiplier equals 1 divided by (1 – 0.6), which is 2.5.

(b) When the *MPC* is 0.4, the multiplier equals 1 divided by (1 – 0.4), which is 1.67.

Nominal national income, real national income and the size of the multiplier

It is important to understand the following relationship:

nominal national income = real national income × average price level

or, in shorthand:

$$Y = Py$$

The size of the multiplier depends on whether we are measuring the *nominal* national income multiplier or the *real* national income multiplier. In our earlier example, the size of the multiplier was 1.25 but this was the nominal national income multiplier. Provided the *SRAS* curve slopes upwards, as is the case in Figure 10.12, the size of the multiplier measured in real terms is always going to be smaller than the nominal national income multiplier. This is because part of the multiplier effect deflects into a rising price level. The growth of real income is restricted to the distance between y_1 and y_2, with the price level rising from P_1 to P_2. Indeed, if the *SRAS* curve is vertical rather than upward sloping, the size of the multiplier measured in real *income* terms would be zero. In this situation, the multiplier effect resulting from an increase in government spending would lead solely to inflation and *not* to rising real output.

TEST YOURSELF 10.19

Explain the difference between nominal national income and real national income.

SECTION 10.4 SUMMARY

- Economic activity centres on the production and consumption of goods and services in the economy, together with the employment of the labour, capital and other inputs that produce output.
- *AD*/*AS* diagrams can be used to illustrate how changes in aggregate demand affect the level of real output.
- *AD*/*AS* diagrams do not directly show the employment levels at each level of real output.
- When real output increases, firms generally have to employ more workers to produce the additional goods and services that the output increase involves. Conversely, when real output falls, less labour is required to produce the smaller amount of goods and services now being produced.
- An increase or decrease in any of the components of aggregate demand, such as government spending, usually leads to a multiplier effect.
- The multiplier measures the relationship between a change in any of the components of aggregate demand (*C*, *I*, *G* and *X* – *M*) and the resulting change in the equilibrium level of national income.

10.5 Determinants of short-run aggregate supply

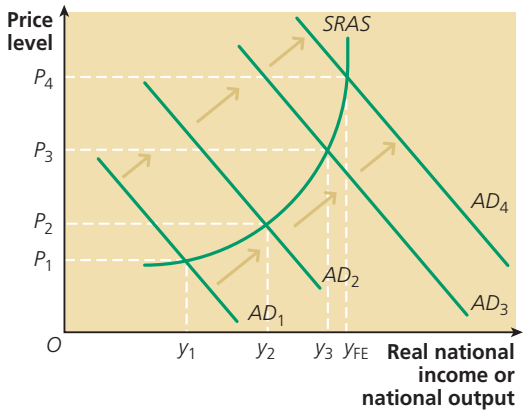


Figure 10.14 How an increase in aggregate demand affects real national income and the price level

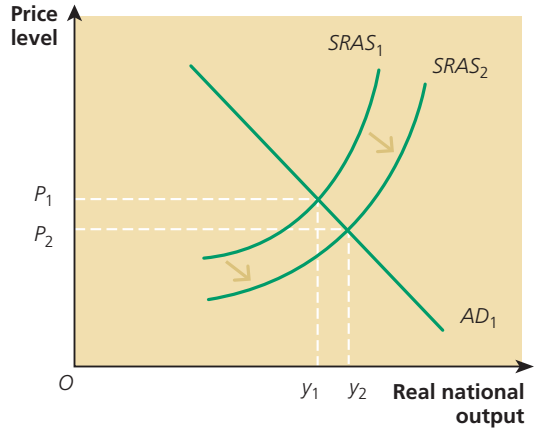


Figure 10.15 A rightward shift of the *SRAS* curve causing a fall in the price level

As mentioned earlier in the chapter, the price level and production costs are the main determinants of the *SRAS* curve. Changes in costs, such as money wage rates, raw material prices, business taxation and productivity, will shift the *SRAS* curve.

All of the *SRAS* curves drawn in this book are 'curved curves' or non-linear curves, which become steeper moving up the curve. This has important implications for monetary and fiscal policy, which are explained further in Chapters 12 and 13. As the *AD* curve shifts to the right along a non-linear *SRAS* curve, such as the one in Figure 10.14, whether or not real income or the price level increases depends on the steepness of slope of the *SRAS* curve.

When the *AD* curve shifts to the right from AD_1 to AD_2, the resulting increase in real output is proportionately greater than the increase in the price level. Real income increases from y_1 to y_2 and the price level rises from P_1 to P_2. This is because the *AD* curve is shifting along the relatively shallow section of the *SRAS* curve. But when the *AD* curve shifts rightwards from AD_3 to AD_4, it is shifting along a much steeper section of the *SRAS* curve. As a result, most of the effect of the increase in aggregate demand falls on the price level rather than on real output. The effect is inflationary rather than reflationary. Indeed, if the *AD* curve were to shift any further to the right beyond AD_4, only the price level, and not real output, would rise. In the diagram, y_{FE} is the full employment level of real income. In this situation, any further increase in aggregate demand results solely in a rising price level, or inflation. At full employment, real income cannot increase, at least in the short run, because the economy is producing at full capacity.

The *slope* of the *SRAS* curve must not be confused with a *shift* of the curve. The *SRAS* curve is constructed under the assumption that all the determinants of aggregate supply *other than the price level* remain unchanged. Should any of these determinants change, the *SRAS* curve shifts to a new position. The curve can shift either to the right (an increase in aggregate supply) or to the left (a decrease in aggregate supply). Figure 10.15 illustrates a rightward shift of the *SRAS* curve, or an increase in aggregate supply, which results in deflation, with the price level falling from P_1 to P_2.

Among the factors that cause a rightward shift of the *SRAS* curve are:

- a fall in businesses' costs of production; these include the costs of imported raw materials and energy
- a fall in unit labour costs, resulting from a fall in wage costs or an increase in labour productivity, the latter possibly caused by better labour training
- a reduction in indirect taxes such as VAT imposed on firms by the government
- an increase in subsidies granted to firms by the government
- technical progress which improves the quality and productivity of capital goods.

A leftward shift of the *SRAS* curve would be caused by the opposite of these factors.

SECTION 10.5 SUMMARY

- The price level and production costs are the main determinants of the *SRAS* curve.
- Changes in costs, such as money wage rates, raw material prices, business taxation and productivity, will shift the *SRAS* curve.
- When the *AD* curve shifts to the right along a non-linear *SRAS* curve, whether or not real income or the price level increases depends on the steepness of the slope of the *SRAS* curve.

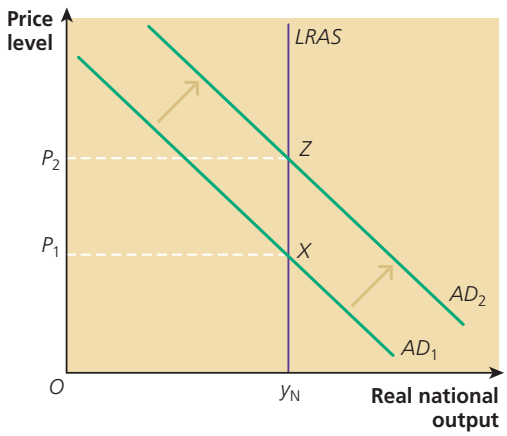


Figure 10.16 An economy's 'normal capacity' level of output

10.6 Determinants of long-run aggregate supply

In the short run, the aggregate supply of real output depends on the average price level in the economy, and the position of the *SRAS* curve is determined by business costs. Other things remaining constant, firms are only prepared to supply more output if the price level rises. However, in the long run, aggregate supply is *not* influenced by the price level. Long-run supply reflects the economy's production potential. The *LRAS* curve is located at the **normal capacity level of output**, which is the level of output at which the full production potential of the economy is being used (see Figure 10.16). To put it another way, it is the maximum level of output that the economy can produce when the economy is on its production possibility frontier.

KEY TERM

normal capacity level of output the level of output at which the full production potential of the economy is being used.

The position of the *LRAS* curve

The position of the *LRAS* curve is determined by the same factors that determine the position of the economy's production possibility frontier. These include the following:

- The state of technical progress.
- The quantities of capital and labour and other factors of production in the economy.
- The mobility of factors of production, particularly labour.
- The productivity of the factors of production, particularly labour productivity.
- People's attitudes to hard work.
- Personal enterprise, particularly among entrepreneurs. The emergence of a large number of risk-taking entrepreneurs is especially important for shifting the *LRAS* curve to the right.
- Related to this, the existence of appropriate economic incentives.
- The institutional structure of the economy. This involves factors such as the rule of law and the efficiency of the banking system. Contract law is particularly important. If contracts are not legally enforceable, normal economic activity can break down, which shifts the position of the *LRAS* curve to the left. And as the financial crisis of 2007 and onwards showed, the inability of banking systems to provide finance to businesses also shifted the *LRAS* curve leftwards. This deepened and lengthened the 'great recession' in 2008 and 2009. Efficient contract law and an efficient banking system 'grease the wheels' of the economy (making it run smoothly), promoting long-run economic growth and an outward shift of the economy's production possibility frontier and a rightward shift of the *LRAS* curve.

SYNOPTIC LINK

We discuss the nature of the banking system in Chapter 12 where we look at the composition of the UK's financial system and various financial markets, along with how the government attempts to keep the system working through monitoring, regulation and support.

QUANTITATIVE SKILLS 10.5

Worked example: interpreting data on the components of aggregate demand

Table 10.3 below contains hypothetical data on the components of aggregate demand in 2023 in an economy experiencing positive economic growth

Table 10.3 Components of aggregate demand, 2023

	£bn at 2020 prices	% change in the year 2022/23
Household consumption	1,048	4.0
Investment	306	5.2
Government spending	292	3.4
Exports of goods and services	496	2.6
Imports of goods and services	570	7.1

Which one of the following statements is supported by the data?

A The economy enjoyed a balance of payments surplus on current account in 2023.
B Investment amounted to approximately 19.5% of aggregate demand in 2023.
C Exports of goods and services was the largest component of aggregate demand in 2023.
D The increase in the value of the components of aggregate demand in 2023 was solely due to inflation.

The only correct answer is statement B. To see that statement B is correct, you should make use of the aggregate demand equation: $AD = C + I + G + (X - M)$. The equation is a reminder that imports of goods and services are a leakage of demand from the circular flow of income. All the other items in the left-hand column in Table 10.3 are positive components of aggregate demand. Adding these up gives the total £2,142 billion, but then the leakage of £570 billion must be subtracted to arrive at the aggregate demand total of £1,572 billion. Finally, the value of gross fixed investment as a percentage of aggregate demand is (£306bn/£1,572bn) × 100, which is approximately 19.5%.

Statements A, C and D are incorrect. Statement A is wrong because there was a current account deficit rather than a surplus in 2023 (£496bn – £570bn), which is a deficit of £74 billion. Consumption was the largest component of aggregate demand in 2023, so statement C is incorrect. Finally, statement D is wrong because economic growth as well as inflation would have contributed to the increase in the value of the components of aggregate demand in 2023.

EXTENSION MATERIAL

A closer look at the 'normal capacity' level of output

Even when producing at 'normal capacity', the economy may still be capable of *temporarily* producing a *higher* level of real output. Figure 10.17 can be used to explain this possibility.

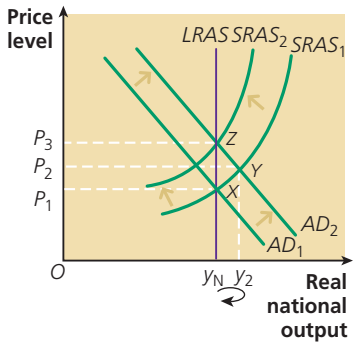


Figure 10.17 The economy can *temporarily* produce a level of real output higher than the 'normal capacity' level of output

We will assume that the economy is initially in macroeconomic equilibrium at point *X* in Figure 10.17, where curve AD_1 intersects both the $SRAS_1$ and the *LRAS* curves. Since the economy is producing on the *LRAS* curve, real output is at its 'normal capacity' level y_N. An increase in aggregate demand, caused perhaps by an expansionary monetary policy (e.g. through lower interest rates), shifts the *AD* curve rightwards to AD_2, which brings about a new short-run macroeconomic equilibrium at point *Y*. In the new equilibrium, the level of real output has risen to y_2, which is above the 'normal capacity' level of output.

However, this level of output, in which the economy is producing above its production potential, cannot be sustained, unless of course the *LRAS* curve itself shifts rightwards. To understand why this is the case, remember that one of the determinants of the position of an *SRAS* curve is the level of money wages in the economy. A higher level of money wages shifts the *SRAS* curve upwards or leftwards, whereas lower money wages would shift the curve rightwards. At Y_2 there are shortages of labour and other factors of production and so excess demand in the labour market and other factor markets leads to rising factor prices. To persuade workers to supply the extra labour needed for y_2 to be produced, given the fact that the price level has risen to P_2, money wages must also rise. As soon as this happens, the *SRAS* curve shifts to $SRAS_2$. As the backward-bending curved arrow lying along the horizontal axis shows, real output falls back from y_2 to the 'normal capacity' level of real output at y_N, located below point *Z* on the diagram.

TEST YOURSELF 10.20

Distinguish normal capacity income from full capacity income.

The Keynesian aggregate supply curve

The vertical *LRAS* curve is sometimes called the free-market *LRAS* curve. This label reflects the view commonly expressed by free-market economists that, provided markets function competitively and efficiently, the economy always operates at or close to normal capacity. In the short run, real output is influenced by the average price level, but in the long run, aggregate supply is determined by maximum normal production capacity, which determines the position of the *LRAS* curve.

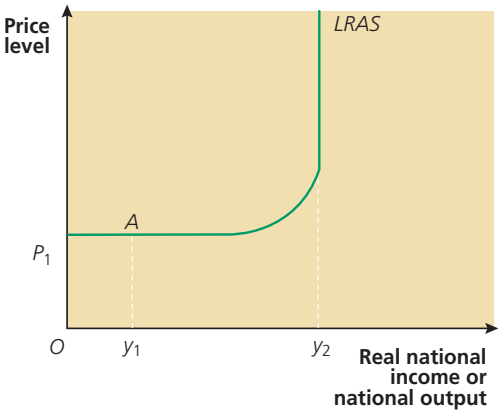


Figure 10.18 The Keynesian 'inverted L-shaped' *LRAS* curve

These days, most economists agree that the *LRAS* curve is vertical. Some economists argue, however, that the *LRAS* curve has a different shape, the 'inverted L-shape' shown in Figure 10.18.

The inverted L-shaped *LRAS* curve is based on the explanation put forward by John Maynard Keynes of the Great Depression in the UK and US economies in the 1930s. Keynes argued that a depressed economy can settle into an under-full-employment equilibrium, shown for example by point *A* on the horizontal section of the *LRAS* curve. At point *A*, the level of real national output is y_1. Keynes believed that without purposeful intervention by the government, an economy could display more or less permanent demand deficiency. Market forces would fail to adjust automatically and achieve full employment. If the government could shift *AD* to the right along the horizontal section of the *LRAS* curve (mainly through expansionary fiscal policy), the existence of significant amounts of spare capacity would lead, in Keynes's view, to a growth in real output (and employment), without an increase in the price level. Eventually, when maximum normal capacity is achieved, the *LRAS* curve becomes vertical for the same reasons that the 'free-market' curve is vertical.

STUDY TIP

Although the Keynesian *AS* curve is often drawn as an 'inverted L' shape, there are variations on this theme. The section of the curve that curves upwards can be extended so that the whole curve looks similar to the *SRAS* curve — but with it becoming vertical at the full-capacity level of output. Sometimes, the section of the Keynesian *AS* curve which is perfectly elastic is often drawn slightly upward sloping — showing that even at lower levels of capacity utilisation, an increase in national output can still mean that a small increase in the price level occurs.

APPLICATION OF ECONOMICS IN THE REAL WORLD 10.3

Keynes on 'Poverty in Plenty'

In 1934, in the depth of the Great Depression, and two years before he published his *General Theory of Employment, Interest and Money*, John Maynard Keynes gave a radio talk in which he expressed his view on the causes of unemployment. Keynes's talk was one of a series, entitled 'Poverty in Plenty', in which a number of economists and public figures gave their views on the title theme. At this time, the UK unemployment rate was 16.7%. Keynes started by summarising the common ground between himself and the other contributors. He then outlined what he saw as the main difference between the various contributors, before arguing that, in his view, the economic system was not self-adjusting. Here is what Keynes had to say.

> We must not regard the conditions of supply, our ability to produce, as the fundamental source of our troubles. It is the conditions of demand which our diagnosis must search and probe for explanation. All the contributors to these talks meet to this extent on common ground. But every one of us has a somewhat different explanation of what is wrong with demand, and, consequently, a different idea of the right remedy.
>
> Though we all start out in the same direction, we soon part company into two main groups. On one side are those who believe that the existing economic system is, in the long run, a self-adjusting system, though with creaks and groans and jerks, and interrupted by time lags, outside interference and mistakes.
>
> On the other side of the gulf are those who reject the idea that the existing economic system is, in any significant sense, self-adjusting. They believe that the failure of effective demand to reach the full potential of supply, in spite of human psychological demand being far from satisfied for the vast majority of individuals, is due to much more fundamental causes.
>
> The strength of the self-adjusting school depends on it having behind it almost the whole body of organised economic thinking and doctrine of the last hundred years. Now, I range myself with the heretics on the other side of the gulf. There is, I am convinced, a fatal flaw in that part of the orthodox reasoning which deals with the theory of what determines the level of effective demand and the volume of aggregate employment. The system is not self-adjusting, and, without purposive direction, it is incapable of translating our actual poverty into our potential plenty.

Follow-up questions

1. What is meant by 'effective aggregate demand'?
2. Suggest why there may be too little effective aggregate demand in an economy.
3. Why do free-market or anti-Keynesian economists believe the economic system is self-adjusting?
4. After the end of recession in 2009, the UK economy grew continuously, albeit at a slow annual rate of growth, until early 2020. Explain two causes of this growth.

TEST YOURSELF 10.21

Write short answers to the following questions.

(a) Is net national income the same as GDP?

(b) What is the difference between nominal and real national income?

(c) What is the difference between saving and investment?

(d) List the components of aggregate demand.

(e) What is deficient aggregate demand?

(f) Why is the *LRAS* curve usually assumed to be vertical?

EXTENSION MATERIAL

AD/AS analysis and the rate of inflation

Students are often puzzled by a particular feature of the *AD/AS* model as taught and learnt at A-level. If you look carefully at conventional *AD/AS* diagrams such as Figure 10.19, you will see that a leftward movement of the *AD* curve results in a *fall* in the price level: that is, deflation. However, we all know that, except in a deep recession or immediately after a very large fall in world commodity prices, contractionary fiscal or monetary policy that shifts the *AD* curve to the left results not in falling prices, but in a slowdown in the rate of inflation, or disinflation. Average prices still rise, but at a slower annual rate of increase.

In Figure 10.20, by contrast, contractionary fiscal or monetary policy, and indeed a fall in any of the components of aggregate demand, reduces the rate of inflation (shown in Figure 10.20 by the symbol $\dot{P}$), but not necessarily the price level.

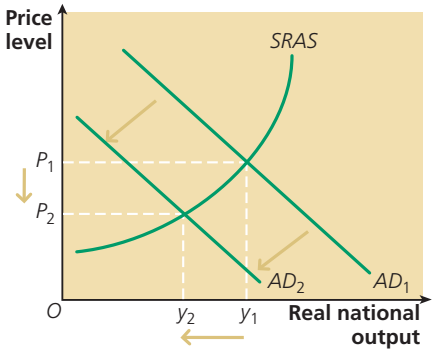


Figure 10.19 The traditional *AD/AS* model

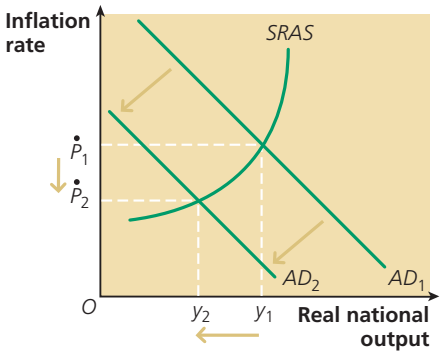


Figure 10.20 *AD/AS* analysis and the rate of inflation

TEST YOURSELF 10.22

Table 10.4 shows selected hypothetical macroeconomic indicators for an economy in 2022 and 2023.

Table 10.4 Macroeconomic indicators, 2022 and 2023

	% change in real GDP	% change in prices	% change in earnings	% unemployment rate
2022	4.2	4.6	9.4	8.4
2023	6.0	4.2	9.2	7.2

Explain why statement B of the following statements is correct, and also why the other three statements are incorrect.

A The price level fell between 2022 and 2023.

B Economic growth was accompanied by a falling unemployment rate.

C A fall in the rate of unemployment was accompanied by a falling price level.

D Earnings fell by 0.2% between 2022 and 2023.

SECTION 10.6 SUMMARY

- It is important to distinguish between a short-run aggregate supply (*SRAS*) curve and a long-run aggregate supply (*LRAS*) curve.
- *SRAS* curves generally slope upwards, but the *LRAS* curve is generally assumed to be vertical with its position determined by the economy's normal capacity level of output.
- The normal capacity level of output is the level of output at which the full production potential of the economy is being used.
- A rightward movement of the *LRAS* curve illustrates long-term economic growth.
- Some Keynesian economists have argued that the *LRAS* curve is horizontal to the left of the full employment level of real output, becoming vertical at the full employment level of output.

Questions

1 Explain the meaning of the circular flow of income.

2 What are the components of aggregate demand? Illustrate how an increase in exports affects the position of the *AD* curve and equilibrium national income.

3 Explain how changes in the rate of interest affect consumption and saving.

4 Distinguish between the accelerator and the multiplier.

5 Explain the shape of the aggregate supply curve, both in the short run and in the long run.

6 Considering only the *SRAS* curve and not the *LRAS* curve, explain how an increase in aggregate demand may affect output and the price level.



11

Economic performance

This chapter returns to a number of topics mentioned briefly in Chapters 9 and 10. It focuses on the core macroeconomic topics of economic growth, unemployment and inflation, explaining and exploring the topics in greater detail. In particular, the *AD/AS* macroeconomic model introduced in Chapter 10 is used to analyse various issues related to each of the topics.

The chapter is about economic performance. A country's macroeconomic performance can be judged by how successful the economy is at achieving the four main objectives of macroeconomic policy outlined in Chapter 9. The first three of these are covered in this chapter — can the economy achieve and then sustain a satisfactory rate of economic growth, relatively full employment, and relative price stability and control of inflation?

The meaning of economic performance having been explained here, Chapters 12 and 13 will go on to investigate how monetary policy and fiscal policy are used to try to improve economic performance. Then, in Chapter 14 we consider the international economy which encompasses the fourth main objective of macroeconomic policy — avoidance of an external deficit on the current account of the balance of payments.

LEARNING OBJECTIVES

These are to understand:

- the demand-side and supply-side determinants of economic growth
- the economic cycle and output gaps
- the benefits and costs of economic growth
- the effects of economic shocks on the economy
- the main types of unemployment
- demand-side and supply-side causes of unemployment
- how global events can impact on UK unemployment and inflation
- demand-pull and cost-push inflation, and the monetarist theory of inflation
- how changes in world commodity prices affect domestic inflation
- the effect of deflation on the economy
- how conflicts that arise from trying to achieve different macroeconomic policy objects affect economic performance in the short run and the long run
- the Phillips curve relationship.

11.1 Economic growth and the economic cycle

Chapter 1 and Chapter 9 have already introduced you to economic growth. Chapter 1 defined economic growth and, with the help of a production possibility frontier, distinguished between short-run and long-run growth.

Short-run and long-run economic growth

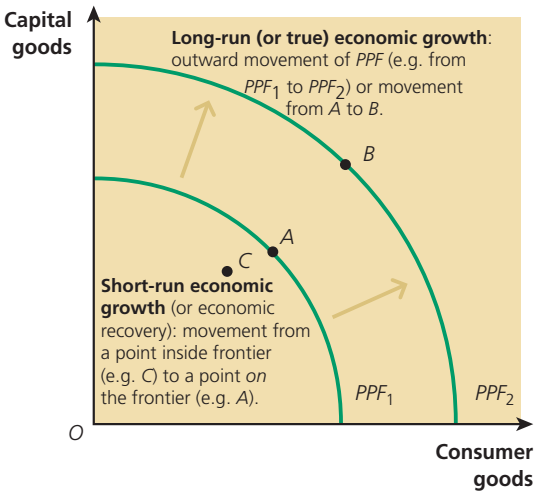


Figure 11.1 Short-run and long-run economic growth illustrated on an economy's production possibility frontier

As we explained in Chapter 1, the definition and the measurement of economic growth are not the same thing. The *definition* of economic growth was first introduced to you in Chapter 1, in the context of the outward movement of the economy's frontier. We defined economic growth as an increase in the *potential* level of real output the economy can produce over a period of time. The *measurement* of economic growth, by contrast, is performed simply by measuring the percentage annual change in real GDP, or possibly real GDP per capita. We then went on to distinguish between short-run and long-run economic growth, both of which are illustrated again in Figure 11.1.

Short-run growth brings idle resources into production and takes up the slack in the economy. By contrast, long-run economic growth, depicted by an outward movement of the economy's production possibility frontier, increases the economy's production potential.

Demand-side and supply-side determinants of short-run growth

KEY TERM
demand side relates to the impact of changes in aggregate demand on the economy; associated with Keynesian economics.

Short-run economic growth, which takes up spare capacity within the economy, is primarily caused by **demand-side** changes in the economy. Anything which shifts the position of the aggregate demand (*AD*) curve is a demand-side influence on the economy. A rightward shift of the *AD* curve along an upward-sloping *SRAS* curve illustrates demand-side causes of short-run economic growth.

A good starting point for explaining demand-induced short-run economic growth is the aggregate demand equation, which we first mentioned in Chapter 10, sections 10.2 and 10.3. To remind you, the aggregate demand equation is:

$$AD = C + I + G + (X - M)$$

where C, I, G and $(X - M)$ are the symbols used for the components of aggregate demand: consumption, investment, government spending and net export demand (exports minus imports). $(X - M)$ is also known as net trade.

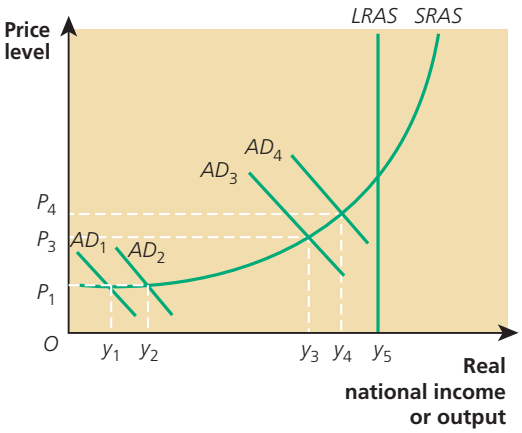


Figure 11.2 Comparing the effects on economic growth of rightward shifts of the *AD* curve

If any of these components of aggregate demand change, the *AD* curve shifts to a new position. An increase in consumption, investment, government spending or net export demand causes the *AD* curve to shift to the right, as shown in Figure 11.2. What then happens in the economy depends on the shape and slope of the aggregate supply curve. Figure 11.2 shows an *SRAS* curve which is horizontal at very low levels of real income, then slopes upwards, and eventually crosses the economy's vertical *LRAS* curve.

If, initially, the level of real output is well below the normal capacity level of output (y_5 in Figure 11.2), a rightward shift of the *AD* curve from AD_1 to AD_2 increases real output, but has no effect on the price level. Real output increases from y_1 to y_2, but the price level stays the same at P_1. This is because the *SRAS* curve is horizontal at very low levels of real output.

However, as output rises, the *SRAS* curve begins to slope upwards, which means that further increases of aggregate demand lead to inflation as well as rising real output. For example, when the *AD* curve shifts from AD_3 to AD_4, the price level rises from P_3 to P_4. And as the aggregate demand curve shifts closer to the normal capacity level of output, short-run economic growth is accompanied by higher rates of inflation. Short-run growth gradually absorbs the spare capacity in the economy until, when y_5 is reached (the normal capacity level of output), there are no longer any idle resources and the economy is producing on the *LRAS* curve. At this point, for growth to continue, short-run growth must give way to long-run growth.

To focus on the **supply-side** determinants of short-run economic growth, anything which shifts the position of the *SRAS* curve to the right will lead to short-run economic growth. This could be a fall in money wage rates or any other cost of production, a fall in taxes imposed on firms or an increase in subsidies paid to firms. However, the main impact of supply-side improvements is on the position of the *LRAS* curve, as we will now explain.

KEY TERM

supply side relates to changes in the potential output of the economy, which is affected by the available factors of production (e.g. changes in the size of the labour force and the productivity of labour).

STUDY TIP

Make sure you understand the difference between aggregate demand and aggregate supply.

- Aggregate demand is determined by factors affecting spending.
- Aggregate supply is determined by factors affecting the costs of production (*SRAS*) and the determinants of an economy's productive capacity (*LRAS*).

TEST YOURSELF 11.1

Explain the difference between short-run and long-run economic growth.

SYNOPTIC LINK

Chapter 10 explains the *AD/AS* macroeconomic model. In this chapter we will frequently make use of this economic model to show the causes of economic growth, unemployment and inflation, as well as how policy changes can help achieve the government's objectives (and any conflicts that may be created by policy changes).

APPLICATION OF ECONOMICS IN THE REAL WORLD 11.1

Recession: 'now you see it, now you don't'

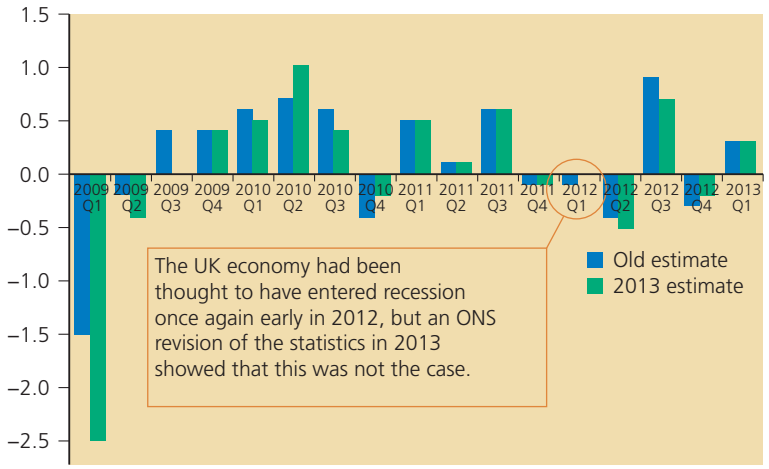


Figure 11.3 GDP, quarter-on-quarter growth, revised figures, 2009–13

The Office for National Statistics (ONS) regularly revises its published statistics. When this happens, the beginning, end and depth of recessions (and of periods of positive growth) can change. Before revised statistics were published in 2013, the official view was that the UK economy had entered the second dip of a 'double-dip' recession. Some economists were even predicting that a 'third dip' was on the cards. The revised statistics — if they can be trusted — show that the 'double dip' did not happen.

The new statistics led to newspaper headlines such as 'The recession that never was: the small change which makes a big difference' in the *Daily Mail*. The data in Figure 11.3, published by the ONS in June 2013, show what 'apparently' happened to quarterly UK economic growth between the second quarter of 2008 and the first quarter of 2013. The graph begins by showing the deep recession which lasted from the second quarter of 2008 until the end of the second quarter of 2009 (five quarters in total).

Figure 11.3 contains two sets of data. The blue bars in the graph show the ONS's estimate of quarterly growth published before June 2013. These data show the 'second dip' of a 'double-dip' recession occurring between the fourth quarter of 2011 and the end of the second quarter of 2012. But according to the revised data, shown by the green bars, there was zero growth in the first quarter of 2012: that is, neither positive nor negative. Hence the 'second dip' never happened and there was no 'double-dip' recession. However, the revised data also show the 2008–09 recession being deeper than previously thought.

Monthly data for GDP are normally published around six weeks after the end of each month. Data for quarterly GDP also appear around six weeks after the end of each quarter. Although GDP data are produced using the output, income and expenditure methods of calculation, the early estimates are based mainly on the output method. This means they are not always accurate due to omissions and errors in the production calculation. Revised data for quarterly GDP are usually published around three months after the end of each quarter — these data incorporate all three methods and are more accurate. This does not mean that the original data serve no purpose. The difference between the original and revised data usually does not vary by a significant amount (in percentage terms) and early availability of data can be very helpful for formulating policy decisions.

A conclusion you may draw from this story is never to trust economic data. The ONS follows the 'continuous revision' method of publishing economic statistics. When new information comes to light, the data are revised.

Incidentally, the UK economy entered recession again more recently as economic growth was negative for the first two quarters of 2020. This was the result of the 'lockdown' of many businesses during the early stages of the Covid-19 pandemic. However, although growth was erratic through the rest of 2020 and in 2021, the UK did not officially enter recession again — the only subsequent quarter of negative economic growth was in the first quarter of 2021. Hence there was no 'double-dip' recession.

Follow-up questions

1. Describe how a recession is defined in the UK and describe three features of a recession.
2. Find out how a recession is defined in other countries, including the USA.
3. Research what has happened to UK economic growth since this book was published.
4. How has the UK growth rate compared to that of the eurozone since the start of 2022 — when most countries 'reopened' in full after the pandemic?

TEST YOURSELF 11.2
'Some economists argue that recessions are necessary for restructuring the economy.' Evaluate this view.

KEY TERM
trend growth rate the rate at which output can grow, on a sustained basis, without putting upward or downward pressure on inflation. It reflects the annual average percentage increase in the productive capacity of the economy.

Demand-side and supply-side determinants of long-run growth or trend growth

The trend growth rate

The economy's **trend growth rate** (or potential growth rate), which is often called the long-run growth rate, is the rate at which output can grow, on a sustained basis, without putting upward or downward pressure on inflation. Before the UK economy entered recession in 2008, the UK's trend growth rate was about 2.5% a year. At first sight, this growth rate appears low, especially when compared to higher trend growth rates in newly industrialising countries (NICs), and especially China. Nevertheless, the UK's trend growth rate was similar to the long-run growth rates of other developed economies in western Europe and North America. The absolute increase in real output delivered by a 2.25% growth rate may also exceed that delivered by a 10% growth rate in a much poorer country. Moreover, because of the compound interest effect, a 2.25% growth rate means that average UK living standards double every generation or so.

It has been estimated that UK trend growth has fallen over the last ten years to below 2%. There is also some evidence emerging that the trend growth for the UK may now be closer to 1%. The number of people who are economically inactive (i.e. of working age but choosing not to work or seek work) has increased in recent years. This is linked to those workers who were 'furloughed' during the Covid-19 pandemic but have since left the labour force. (The furlough scheme was a system in operation between 2020 and 2021 where the government paid businesses to keep workers employed by, in effect, paying their wages — or a significant proportion of their wages.) This group of people may have chosen to retire early or simply not to work — which may be achievable if they have other earners in their household.

The causes of long-run economic growth

We have already seen that short-run economic growth is caused by an increase in aggregate demand which makes use of spare capacity and takes the economy from a point inside the economy's production possibility frontier to a point on the frontier. By contrast, the main causes of long-run growth lie in the supply side of the economy.

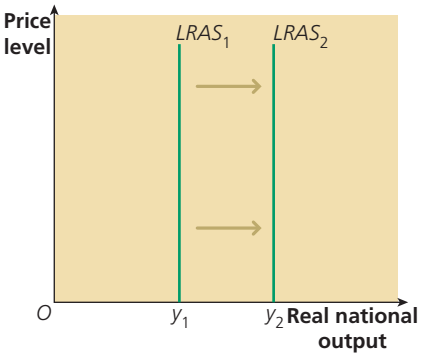


Figure 11.4 Long-run economic growth illustrated by a rightward shift of the *LRAS* curve

In Figure 11.4, once full capacity is reached at the level of real output y_1 (the initial normal capacity level of output), for sustainable economic growth to continue the *LRAS* curve must shift rightwards from $LRAS_1$ to $LRAS_2$. The new full capacity level of real output is y_2. As Chapter 10 mentioned, the determinants of long-run growth include improvements in technology resulting from investment and technical progress, increases in the size of the labour force, together with improvements in productivity, attitudes, enterprise, the mobility of factors of production and the economic incentives faced by entrepreneurs and the workers they employ.

To state it another way, the determinants of long-run economic growth relate to the economy's supply side. Supply-side reforms which increase labour productivity, bringing about higher output per worker, result from the investment decisions made by firms in pursuit of profit. These can be backed up and reinforced by successful supply-side policies undertaken by government.

However, aggregate demand still plays a role in bringing about long-run economic growth. This is because firms will only produce more goods and services if there is sufficient demand in the economy to absorb the extra output. Aggregate demand must increase to match the increase in aggregate supply.

SYNOPTIC LINK

The roles of supply-side reforms and supply-side policies are examined in more detail in Chapter 13 on fiscal policy and supply-side policies.

EXTENSION MATERIAL

Economic growth theories

As explained above, long-run economic growth is caused by both supply-side and demand-side factors. The former is the prime mover, but sufficient aggregate demand has to be generated to absorb the extra output produced by the growth process. The immediate supply-side cause of long-run growth is *increased labour productivity*, which itself results from investment in, and accumulation of, capital goods and human capital, and from technological progress.

There are two main theories of long-term economic growth: neoclassical growth theory and new growth theory.

Neoclassical growth theory

The older theory, known as neoclassical growth theory, which was developed by Professor Robert Solow in the 1950s, argues that a sustained increase in investment increases the economy's growth rate, but only temporarily. The ratio of capital to labour goes up, the marginal product of capital declines and the economy moves back to a long-term path, which is determined by output growing at the same rate as the workforce, plus a factor to reflect improving labour productivity.

In neoclassical growth theory, the rate at which labour productivity improves is determined by technological progress. But the theory does not explain why technological progress occurs. Technological progress is assumed to occur outside of the model. This is the theory's weakness. The causes or determinants of the ultimate engine of economic growth, namely technological progress, are exogenous to the theory (exogenous means 'outside'). Neoclassical growth theory therefore fails to provide a complete explanation of economic growth.

New growth theory

In recent years, neoclassical growth theory has been replaced to a significant extent by new growth theory. New growth theory is also called *endogenous growth theory*, reflecting the fact that, unlike in neoclassical growth theory, the determinants of technological progress are brought inside the theory. The three main sources of technological progress explained by new growth theory are profit-seeking research, openness to foreign ideas and accumulation of human capital:

- **Profit-seeking research.** The rate at which technological progress occurs depends on the stock of ideas. The flow of new ideas thought up by current researchers adds to the 'capital stock' of existing ideas. How many new ideas there are depends on the number of researchers, but the extent to which new ideas improve technological progress depends on whether or not 'over-fishing' occurs. 'Over-fishing' means that the discovery of new ideas makes it harder to find further new ideas. But conversely, the opposite may be true, as accumulating ideas may make researchers more rather than less productive. This is the 'standing on the shoulders' effect. Paul Romer, one of the most influential developers of new growth theory, assumes that this effect is dominant, which means that countries with more researchers can have higher growth rates.
- **Openness to foreign ideas.** Economic growth can derive either from domestic innovation or from technological transfer from other countries. In 1999, Cameron, Proudman and Redding argued that the rate at which technological progress occurs in a country depends on three factors: the domestic rate of growth of technology in the absence of technology transfer; the rate at which technology can be adopted from abroad; and the proportion of foreign technologies that can be adopted. This means that for a technology follower, technology grows at its own domestic rate of technology growth, plus some extra 'catch-up' generated by technology-leading countries.
- **Accumulation of human capital.** Human capital accumulates through educating and training a skilled workforce from among a country's indigenous population, and through migration from other countries. Migration adds to human capital provided migrants possess appropriate education and skills, or are willing and able to attain them. A high level of human capital is best regarded as a necessary but not a sufficient condition for successful economic growth. This is because technological change requires workers to possess the skills and aptitudes required for adapting to new technologies, rather than those that used to be necessary for old, declining technologies.

New growth theory suggests that appropriate government intervention can create the supply-side conditions which favour growth. These include:

- conditions that encourage profit-seeking research and appropriate accumulation of human capital
- externalities or external economies provided by the state which benefit private-sector businesses
- patent legislation and a judicial system that enforces the law of contract and intellectual property rights, which create the incentive for firms to innovate.

The costs and benefits of economic growth

For many people and most economists, achieving a satisfactory and sustained rate of economic growth is arguably the most important of all the macroeconomic objectives that governments wish to achieve. Without growth, other objectives, particularly full employment and competitive export industries, may be impossible to attain. And when growth becomes negative, as in the recession that started in the UK in 2008, people become all too aware of the rapid disappearance of the fruits of growth. For most people, standards of living fall, with the most unfortunate losing their jobs as the industries that used to employ them collapse or slim down.

However, in the long run, high rates of economic growth may not be sustainable, especially in developing economies. This is particularly the case in the 'emerging-market' countries (countries that are developing but are beginning to engage with the global economy), which have recently been growing at a far faster rate than developed economies. Here the rapid using-up of finite resources and the pollution and global warming that spin off from economic growth will result increasingly in desertification, water shortages, declining crop yields, famines and wars.

Some economists argue an opposite effect, namely that one of the benefits of growth, at least as far as advanced developed economies are concerned, is the development of environmentally friendly technologies. These reduce the ratio of energy consumption to GDP. Nevertheless, rich developed economies, especially the USA, continue to be, at least for the time being, among the world's biggest consumers of energy and the biggest polluters.

Economic growth can have a number of costs that reduce economic welfare. Some of these, followed by some of the benefits of growth, are listed below.

Costs of economic growth

- Economic growth uses up finite resources such as oil and minerals that cannot be replaced.
- Economic growth leads to pollution and other forms of environmental degradation, with the Earth eventually reaching a tipping point, beyond which it cannot recover.
- Growth can destroy local cultures and communities, and widen inequalities in the distribution of income and wealth.
- Economic growth leads to urbanisation and the spread of huge cities, which swallow up good agricultural land.
- In its early phases, economic growth leads to a rapid growth in population, more mouths to feed, and more people who are poor.
- Growth produces losers as well as winners. Countries suffering low growth may enter a vicious circle of declining business confidence, low profits, low investment, a lack of international competitiveness, even lower profits, zero growth and so on.



Air pollution in New Delhi, India, shows one cost of economic growth

Benefits of economic growth

- Economic growth increases standards of living and people's welfare.
- Growth may lead to communities that work to improve the environment.
- Growth provides new and more environmentally friendly technologies.
- Economic growth has increased the length of people's lives and has provided the means to reduce disease.
- Economic growth provides a route out of poverty for much of the world's population.
- Economic growth produces a 'fiscal dividend', namely the tax revenues that growth generates. Tax revenues can be used to correct market failures and to provide infrastructure, thereby increasing the economic welfare of the whole community.
- For a particular country, economic growth can generate a 'virtuous circle' of greater business confidence, increased investment in state-of-the-art technology, greater international competitiveness, higher profits, even more growth and so on.

SYNOPTIC LINK

Economic growth and economic development obviously overlap. Chapter 15, section 15.2, investigates economic development further. The chapter also describes alternatives to national income as measures of economic development, such as the United Nations Human Development Index (HDI).

The impact of growth on individuals, the economy and the environment

Some individuals, mostly the rich but also many of the less well-off, obviously benefit from the higher welfare and incomes, and the better prospects for the future that growth brings about. But others, including many of the poor, can suffer from loss of jobs and increased inequality that can also result from economic growth. For example, increased automation brought about by faster economic growth may reduce overall demand for labour.

Economic growth can increase the international competitiveness of the economy, which implies that other less competitive economies will not enjoy so many of the fruits of rapid growth. The reverse, however, is true if other countries grow at a faster rate.

With regard to the environment, economic growth may have more costs than benefits, leading to an outcome in which growth becomes unsustainable. Growth may consume more and more finite resources, leading to resource depletion, and it also leads to resource degradation through pollution and loss of biodiversity.

STUDY TIP

Don't confuse *sustainable growth* with *sustained growth*. The latter simply means achieving a particular rate of economic growth (on average) over a number of years. For example, from the end of recession in 1992 until the onset of the next recession in 2008, the UK sustained an average annual growth rate of about 2.75% over this period. However, a growth rate which is sustained over quite a long period may nevertheless be unsustainable if it leads to environmental problems, the depletion of vital finite resources and the degradation of others.

SYNOPTIC LINK

See the section on environmental market failures in Chapter 8 (pages 241–44), where we consider how the lack of established property rights may directly lead to over-use and depletion of natural resources — perhaps more likely in a developing economy.

EXTENSION MATERIAL

The environment and the sustainability of economic growth

In 1973 the first oil crisis, in which the price of crude oil more than doubled, led to British people becoming aware of how the environment — in the guise of raw material and energy shortages — might severely limit a government's ability to achieve high employment levels and continuing economic growth. Fuel shortages and increased energy prices drew people's attention to two publications of the then new but fast-growing ecology or environmentalist movement. The first of these was a document called 'A blueprint for survival', published in the January 1972 edition of the British journal *The Ecologist*, which was followed a few months later by the publication in New York of a vastly influential book, *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*.

The authors of 'A blueprint for survival' wrote:

> The principal defect of the industrial way of life with its ethos of expansion is that it is not sustainable. Its termination within the lifetime of someone born today is inevitable — unless it continues to be sustained for a while longer by an entrenched minority at the cost of imposing great suffering on the rest of mankind. Our task is to create a society which is sustainable and which will give the fullest possible satisfaction to its members. Such a society by definition would not depend on expansion but on stability.

The Limits to Growth, published a few months later in 1972, suggested that humanity would soon face a threefold dilemma: the impending exhaustion of the world's non-renewable natural resources; the world's pollution problem becoming so severe that the capacity for regeneration becomes exhausted; and as population continues to grow, a point being reached where humankind destroys itself through sheer weight of numbers. Commenting on *The Limits to Growth*, one reviewer wrote:

> Man is heading for disaster if, overall, mankind does not learn to limit economic growth. That such a thesis would be unpopular with economists was to be expected. All their training, their research and their ethos has been concerned with ways in which the economy can be stimulated and expanded. Practically no economist has given thought to the problems of economic stability. For one thing stability is read as stagnation, and for another it is hard for an economist to find anyone to employ him who is interested in economic equilibrium. It has no appeal to governments, board chairmen, company presidents or even to used-car salesmen.

The authors of 'A blueprint for survival' and *The Limits to Growth* based many of their forecasts and predictions on the assumption of accelerating rates of population growth and resource use. This is known as exponential growth. Figure 11.5 illustrates the forecast in 'A blueprint for survival' of exponential growth in demand for crude oil or petroleum and exponential decline in petroleum reserves. The dashed lines show the total world oil reserves that, in 1975, the authors of the 'Blueprint for survival' estimated would be available for extraction.

The forecasts of resource use and depletion made in *The Limits to Growth* were rather more sophisticated,

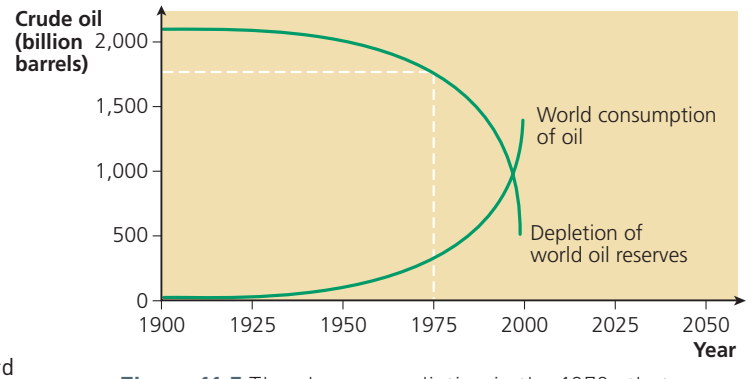


Figure 11.5 The gloomy prediction in the 1970s that exponential growth of oil consumption, matched by exponential depletion of known oil reserves, would lead to the world running out of oil by 2000

but they were still based on the assumption of exponential rates of growth of use and depletion. In their most optimistic scenario, the authors of *The Limits to Growth* argued that it might be possible to achieve a state of equilibrium, provided restrictions were placed on population growth and industrialisation — the two elements limiting growth which they argued principally cause resource exhaustion and pollution. However, increases in the technology to extract oil have meant that the likely date of depletion is further away. Increased use of renewable energy sources may also mean that the supply of oil and other non-renewable sources does not run out.

The economic cycle

Fluctuations in economic activity occur in two main ways: through seasonal fluctuations and through cyclical fluctuations taking place over a number of years. Seasonal fluctuations are largely caused by changes in the weather. Examples include the effect of very cold winters closing down the building trade and seasonal employment in travel and tourism.

Upswings and downswings in economic activity which are longer than seasonal fluctuations are called the **economic cycle**. Economic cycles (also known as **business cycles** and **trade cycles**), which can be between approximately 4 and 12 years long, are caused primarily by fluctuations in aggregate demand. In recent years, supply-side factors such as supply shocks hitting the economy have also been recognised as causes of economic cycles.

Figure 11.6 shows two complete economic cycles, together with a line giving the economy's trend output, from which the economy's long-term growth rate

TEST YOURSELF 11.3
What is the difference between sustained economic growth and sustainable economic growth?

KEY TERM
economic cycle (also known as a **business cycle** or **trade cycle**) upswing and downswing in aggregate economic activity taking place over 4 to 12 years.

KEY TERM

actual output level of real output produced in the economy in a particular year — not to be confused with the trend level of output, which is what the economy is capable of producing when working at full capacity. Actual output differs from the trend level of output when there are output gaps.

can be calculated. **Actual output** rises and sometimes falls in the different phases of the economic cycle. Short-term growth, measured by the percentage change in real GDP over a 12-month period, also varies in the different phases of the economic cycle. In the cycle's upswing, growth is positive but, as Figure 11.6 shows, 'growth' becomes negative if and when a recession occurs in the cyclical downturn.

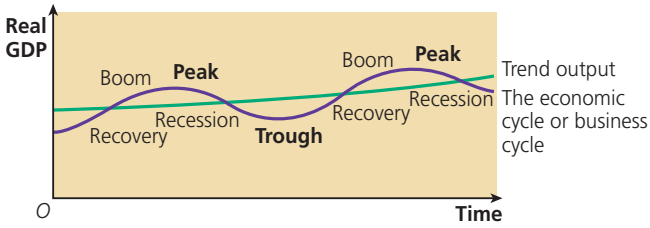


Figure 11.6 The phases of the economic cycle and the trend output line

The phases of the economic cycle shown in Figure 11.6 are recovery, boom and the recessionary phase. Recovery occurs when real GDP begins to grow after the end of a recession, or at the end of the trough of the economic cycle. With continuing short-run growth in real output, recovery gives way to the boom phase of the economic cycle when the level of real output becomes greater than the trend level of output. The boom ends when the upswing of the economic cycle gives way to the next downswing, which is shown in Figure 11.6 by the economy entering a recession. (A severe recession lasting two years or more is sometimes called a depression. Unlike a recession, defined in the UK as real national output falling for six months or more, there is no generally accepted definition of a depression.)

In contrast to Figure 11.6, the next diagram, Figure 11.7, shows continuous *positive* economic growth with an absence of *negative* growth. In the downswings shown on the diagram, the rate of positive growth slows down, but negative growth does not occur. For example, between 2010 and 2014, UK economic growth was never negative but it was generally slow and below increases in the level of trend output.

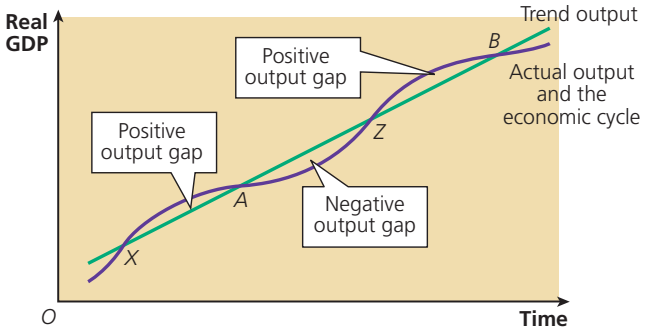


Figure 11.7 Upswings and downswings in the economic cycle, but no recessions

Figure 11.7 illustrates what happened to economic growth in the years before the 2008–09 recession, and also in the recovery period after the recession, though in the latter period the underlying rate of growth was less than had been the case before the recession. During the recession itself, real output fell.

Figure 11.8 gives a better representation of what happened to economic growth before, after and during the 2008–09 recession. The graph shows that, unlike in the previous two recessions, which occurred at the beginning of the 1980s and the 1990s, the 2008–09 recession was much deeper, with real output (shown by the blue line) only recovering to its pre-recessionary level in 2014 and remaining, even then, well below what would have been the level of output had there been no recession. Indeed, the recession may have shifted the economy onto a new trend rate of growth much lower than the pre-recessionary trend growth line.

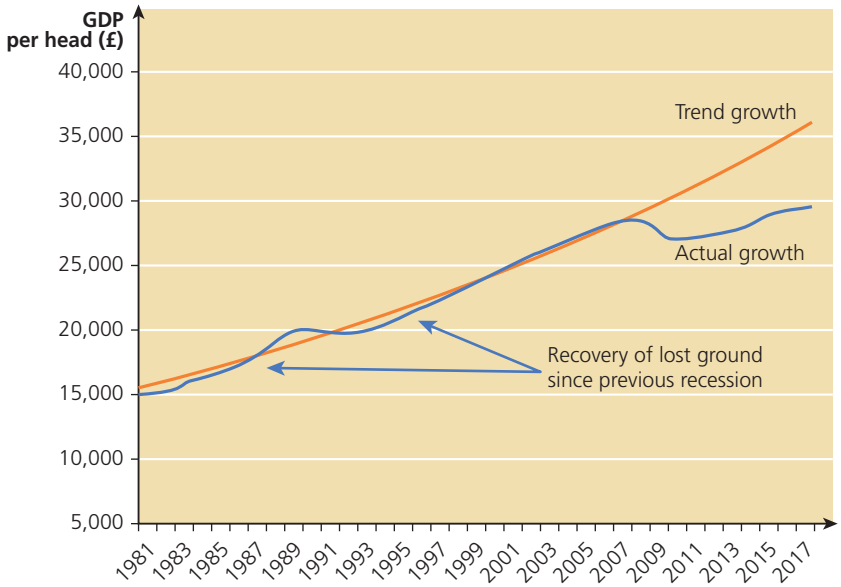


Figure 11.8 Was there a change in UK trend growth after the 2008–09 recession?

The causes of change in the phases of the economic cycle

The principal explanations of the phases of the economic cycle are:

- **Fluctuations in aggregate demand.** In the 1930s John Maynard Keynes argued that economic recessions are caused by fluctuations in aggregate demand, which are caused by consumer and business confidence giving way to pessimism, and vice versa. This is still the prevailing view today.
- **Supply-side factors.** It is now recognised that supply-side factors can also trigger economic cycles. Edward Prescott and Finn Kydland, the 2004 Nobel prize winners in economics, have developed a theory of 'real business cycles', which argues that changes in technology on the supply side of the economy might be as important as changes in aggregate demand in explaining economic cycles.

Other factors that may cause or contribute to cyclical changes include:

- **The role of 'speculative bubbles'.** Rapid economic growth leads to a rapid rise and 'speculative bubble' in asset prices. When people realise that house prices and/or share prices have risen far above the assets' real values, asset selling replaces asset buying. This causes the speculative bubble to burst, which in turn destroys consumer and/or business confidence. People stop spending and the economy falls into recession. The resulting cyclical instability is made worse by the excessive growth in credit and levels of debt, and 'animal spirits' and 'herding'. These terms are used to describe how, in financial markets, share prices and asset prices tend to 'overshoot' when traders suffer a bout of 'irrational exuberance', and then to 'undershoot' when the 'pricking' of the speculative bubble causes asset prices to collapse.
- **Political business cycle theory.** In democratic countries, general elections usually have to take place every four or five years. As an election approaches, the political party in power may attempt to influence the

electorate by engineering a pre-election boom. After the election, the party in power deflates aggregate demand to prevent the economy from overheating, but when the next general election approaches, demand is once again expanded.

- **Outside shocks hitting the economy.** Economic shocks, which were explained in Chapter 10, divide into 'demand shocks', which affect aggregate demand, and 'supply shocks', which impact on aggregate supply. In some cases, an outside shock hitting the economy may affect both aggregate demand and aggregate supply. A commonly quoted example is the effect on other countries of a war in the Middle East. The war might not only affect business confidence in a country such as the UK (a demand shock), but also lead to an oil shortage which increases businesses' costs of production (a supply shock). The 2022 war in Ukraine is seen as a supply shock — with the very large rises in gas prices, adding to the already high cost of living.
- **Changes in inventories.** Besides investing in fixed capital, firms also invest in stocks (inventories) of raw materials, and in stocks of finished goods waiting to be sold. This type of investment is called inventory investment or stock building. Although this accounts for less than 1% of GDP in a typical year, swings in inventories are often the single most important determinant of recessions. Firms hold stocks of raw materials and finished goods in order to smooth production to cope with swings in demand. However, paradoxically, changes in these inventories tend to trigger and exacerbate economic cycles. Stocks of unsold finished goods build up when firms over-estimate demand for finished goods. As the stocks accumulate, firms are forced to cut production by more than the original fall in demand. The resultant destocking turns a slowdown into a recession. In the USA, swings in inventory investment account for about half of the fall in GDP in recent recessions. Destocking has also made UK recessions worse.
- **The Marxist explanation.** Marxist economists explain economic cycles as part of a restructuring process that increases the rate of profit in capitalist economies. Under normal production conditions, a fall in the rate of profit caused by competitive pressure threatens to bankrupt weaker capitalist firms. Marxists believe that recessions create conditions in which stronger firms either take over weaker competitors, or buy at rock-bottom prices the assets of rivals that have been forced out of business. Either way, restructuring by takeover or bankruptcy means that the 'fittest' capitalist firms survive. In Marxist analysis, economic cycles are deemed necessary for the regeneration and survival of capitalism.
- **Multiplier/accelerator interaction.** Keynesian economists have argued that business cycles may be caused by the interaction of two dynamic processes: the multiplier process, through which an increase in investment leads to multiple increases in national income; and the accelerator, through which the increase in income induces a change in the level of investment. Thus the relationship between investment and income is one of mutual interaction: investment affects income (via the investment multiplier), which in turn affects investment demand (via the accelerator process), and in this process income and employment fluctuate in a cyclical manner.
- **Climatic cycles.** Stanley Jevons was one of the first economists to recognise the economic cycle in the nineteenth century. Jevons believed a connection exists between the timing of economic crises and the solar cycle. Variations in sunspots affect the power of the sun's rays, influencing the quality

of harvests and thus the price of grain, which, in turn, affects business confidence and gives rise to trade cycles. Although Jevons's sunspot theory was never widely accepted, the effect named El Niño, which figured in Application in Economics in the Real World 10.1 in the previous chapter, has renewed interest in Jevons's theory.

TEST YOURSELF 11.4

An economy had nominal GDP growth of 8% in the last year, an inflation rate of 5.5% and population growth of 2.5%. What was the approximate change in real GDP per capita over the year?

Indicators, besides changing GDP, which are used to identify phases of the economic cycle

Economic cycles are usually defined and measured in terms of changes in real output or real GDP. However, changes in other variables, including the rate of inflation, investment and unemployment, are also used to describe the phases of the economic cycle. In the upswing of the cycle, especially in the boom phase, excess aggregate demand begins to pull up the price level, a situation exacerbated by cost pressures when labour shortages lead to nominal wages and salaries rising faster than real wages. The reverse happens in the downswing. Business confidence is also higher in the upswing, leading to an increase in private-sector investment. Conversely, a collapse of confidence in the cycle's downswing leads to investment projects being cancelled or postponed.

Employment increases (and unemployment decreases) in the recovery and boom phases of the cycle, though, as noted, eventual labour shortages may lead to a higher rate of inflation. In past UK economic cycles, though less so recently, the employment cycle has tended to lag behind the output or GDP cycle. As people have votes, political parties tend to be more influenced by what is happening to jobs than to what is happening to GDP. The fact that the two cycles are out of phase, as illustrated in Figure 11.9, can lead to governments expanding the economy to reduce unemployment, even when real GDP began its recovery several months beforehand. This can increase inflation. Likewise, governments have sometimes taken demand out of the economy to try to temper a 'frothy' jobs market, even though real GDP has already begun to decline.

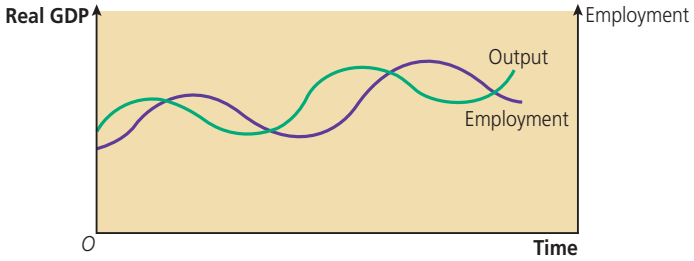


Figure 11.9 The output and employment cycles

The economic cycle and output gaps

KEY TERMS

output gaps show the level of *actual* real output in the economy either higher or lower than the *trend* output level.

positive output gap the level of *actual* real output in the economy is greater than the *trend* output level.

negative output gap the level of *actual* real output in the economy is lower than the *trend* output level.

Figure 11.10 illustrates **output gaps**, a feature of the economic cycle previously shown in Figure 11.7. If the economy's *actual* level of output were always to equal *trend* output, there would be no output gap. (In this situation, there would also be no economic cycles.) Output gaps occur when the level of *actual* real output in the economy is greater or lower than the *trend* output level at a particular point in time. When actual output is above trend output, there is a **positive output gap**. Similarly, when actual output is below trend output, there is a **negative output gap**. (Positive and negative output gaps are both illustrated in Figures 11.7 and 11.10, where the levels of actual output that the economy is producing are above and below the trend output line.) The UK government defines economic cycles as beginning and ending when there are no output gaps, whether or not a recession actually occurs. Thus, in Figure 11.7, an economic cycle beginning at point *X* would end at point *Z*. (Alternatively, points *A* and *B* could be used to mark the beginning and end of an economic cycle.)

TEST YOURSELF 11.5

What is another name for the economic cycle?

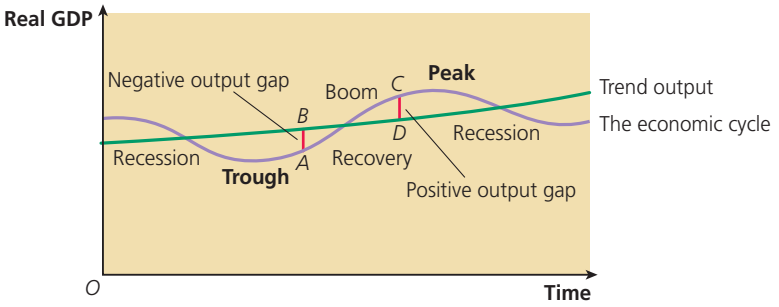


Figure 11.10 Negative and positive output gaps illustrated on an economic cycle diagram

STUDY TIP

Don't confuse an *output gap* with a *productivity gap*, which is the gap between the levels of labour productivity in two countries. Make sure you understand the difference between positive and negative output gaps, as they require different policy approaches to correct output gaps that become too wide.

Output gaps and *AD/AS* gaps

Figure 11.11 shows negative and positive gaps on an *AD/AS* diagram. The economy suffers a negative output gap whenever the level of output is to the left of the *LRAS* curve and *below* the 'normal capacity' level of output y_N. Positive output gaps occur when the economy temporarily produces at a point outside its current production possibility frontier. However, because this represents over-use of capacity, such a point cannot be sustained for long. In the context of the *AD/AS* model, this means that the economy temporarily produces a level of output to the right of the *LRAS* curve, and above the 'normal capacity' level of output y_N.

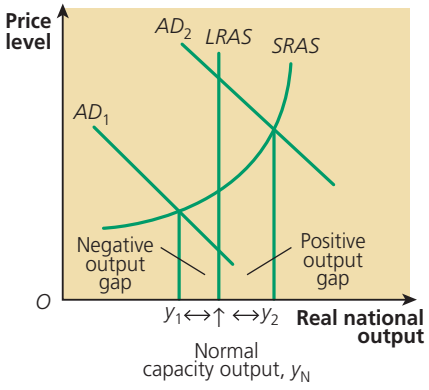
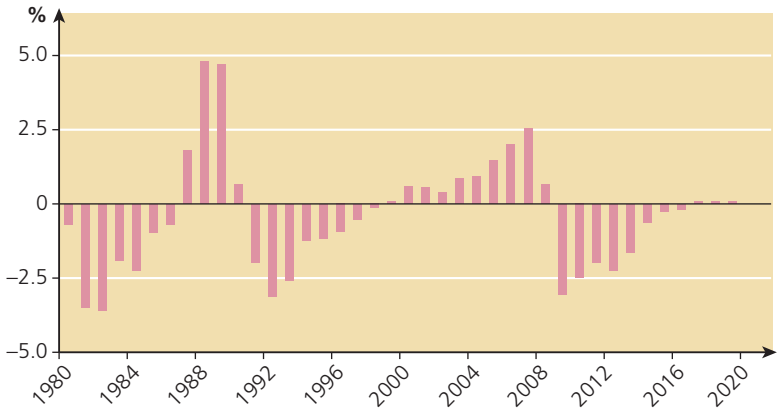


Figure 11.11 Negative and positive output gaps illustrated on an *AD/AS* diagram

Output gaps and the UK economy

Figure 11.12 illustrates how the UK economy performed between 2002 and 2019. When the vertical bars in the graph are positive, the UK experienced a positive output gap. When they are negative, the economy suffered a negative output gap.



Source: IMF

Figure 11.12 UK's output gap measured as a percentage of potential GDP, 1980–2019

In the years between 2000 and 2008 a positive output gap accompanied a booming economy. This was brought to a shuddering halt in the 2008–09 recession. In the subsequent economic recovery, the negative output gap gradually became smaller, until the International Monetary Fund (IMF) forecast a very small positive output gap in 2018. However, this does not indicate a return to economic boom. For many people, particularly those on low incomes and those who live in the more depressed parts of the economy, despite slow positive economic growth, real personal incomes have fallen in many of the years since recovery began. For low-paid workers, economic recovery did not improve their standards of living. The IMF's forecasts for future years after 2018 has hovered around an output gap of approximately zero, with the forecast for 2019 of a very small output gap. Certainly, for 2020 and 2021 the data on output gaps would be very hard to determine. This is because the size of any potential output would have been severely affected by restrictions placed on businesses and output by the UK government as part of its Covid-19 pandemic response.

SECTION 11.1 SUMMARY

- Economic performance reflects success in achieving economic growth, low unemployment and stable prices.
- Economic growth is measured as a change in real GDP rather than as a change in nominal GDP.
- Investment and technical progress are two of the causes of economic growth.
- The economic cycle or business cycle reflects actual output fluctuating around trend output.
- Economic cycles are often caused by changes in aggregate demand, though supply-side factors can also cause fluctuations in economic activity.
- An output gap is the difference between actual output and the trend level of output.
- The UK's trend rate of growth may have changed during and after the 2008–09 recession.

11.2 Employment and unemployment

The meaning and measurement of unemployment

As with economic growth, we have already briefly discussed the meanings of employment and unemployment, together with the measurement of unemployment in Chapter 9. The main point was that the UK government has generally dropped the claimant count as a measure of unemployment, focusing instead on the Labour Force Survey (LFS) measure.

We explained that full employment does not necessarily mean that every member of the working population has a job. Rather, it means a situation in which the number of people wishing to work at the going market real-wage rate equals the number of workers whom employers wish to hire at this real-wage rate.

However, even this definition needs qualifying since, in a dynamic economy, change is constantly taking place, with some industries declining and others growing. This leads to two types of unemployment known as frictional and structural unemployment. We will now look at these and other types of unemployment.

STUDY TIP

Make sure you remember how unemployment is measured in the UK. You must be able to interpret claimant count and Labour Force Survey data.

Voluntary and involuntary unemployment

In the next section, when we explain the different types and causes of unemployment, we will discuss cyclical unemployment, or Keynesian unemployment as it is sometimes called. Writing in the 1930s, at the time of the Great Depression, Keynes argued that deficient aggregate demand in the economy results in **involuntary unemployment**. This is a situation in which the unemployed without jobs wish to work, but there are no jobs available because there is no demand for their labour. The unemployed are willing, able and available to take up employment in any suitable job at the going wage rate, but they cannot find employment. In this situation, unemployment is caused by deficient aggregate demand.

Voluntary unemployment, by contrast, refers to a situation in which workers without jobs are unwilling to take up employment at the going wage rate, often choosing instead to receive benefits paid by the state rather than to seek employment. Although it may seem that in this case unemployed people are lazy, this is not really so. The idea of unemployment being 'voluntary' can be seen as subjective. For example, if someone chooses not to relocate so as to find a job, their unemployment might be considered 'voluntary' by some and 'involuntary' by others.

KEY TERMS

involuntary unemployment occurs when workers are willing to work at current market wage rates but there are no jobs available.

voluntary unemployment occurs when workers choose to remain unemployed and refuse job offers at current market wage rates.

The types and causes of unemployment

The four types of unemployment we will now explain are frictional, structural, cyclical and finally seasonal unemployment. The last of these is the least significant of the four, except of course for many of the workers who become seasonally unemployed.

Frictional unemployment

Frictional unemployment, also known as **transitional unemployment**, is 'between jobs' unemployment. As its name suggests, this type of unemployment results from 'frictions' in the labour market which create a delay, or time lag, during which a worker is unemployed when moving from one job to another. Note that the definition of frictional unemployment assumes that a job vacancy exists and that friction in the job market, caused by the immobility of labour, prevents an unemployed worker from filling the vacancy. It follows that the number of unfilled job vacancies in the economy can be used as a measure of the level of frictional unemployment. Frictional unemployment is usually short term; if it persists, it becomes structural unemployment.

KEY TERM
frictional unemployment (also known as **transitional unemployment**) unemployment that is usually short term and occurs when a worker switches between jobs.

Among the causes of frictional unemployment are geographical and occupational immobilities of labour, which prevent workers who are laid off from immediately filling job vacancies.

The geographical immobility of labour is caused by factors such as family ties and local friendships discouraging people from moving to other parts of the country, ignorance about whether job vacancies exist in other parts of the country, and above all, the cost of moving and difficulties of obtaining housing.

The occupational immobility of labour results from difficulties in training for jobs that require different skills, the effects of restrictive practices such as a requirement that new workers must possess unnecessary qualifications, and race, gender and age discrimination in labour markets.

SYNOPTIC LINK
Occupational and geographical immobility of labour were first mentioned in Chapter 6, section 6.4, as a way in which labour markets are not perfect markets. This is where a market failure in a 'micro' sense has a 'macro' impact.

Redundant workers, who have been laid off by their previous employers because they are no longer needed, may take time to find the types of work they want at wage rates they are prepared to accept. They remain unemployed while involved in job searching. Likewise, young workers joining the labour market for the first time, such as university graduates, may be frictionally unemployed for a short time.

Imperfect information in the labour market, through which the jobless are unaware of jobs that are in fact available, tends to worsen frictional unemployment, as do incentive problems resulting from the fact that people may believe their pay will be swallowed up in taxes and loss of benefits.

EXTENSION MATERIAL

The search theory of unemployment

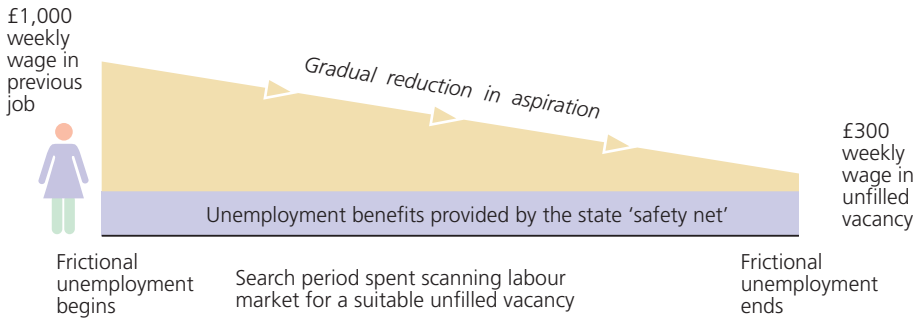


Figure 11.13 Search theory and frictional unemployment

The search theory of unemployment also helps to explain frictional unemployment. Consider the situation illustrated in Figure 11.13. A worker earning £1,000 a week in a skilled professional occupation loses her job. Although no vacancies apparently exist at present in her current line of work, there are plenty of vacancies for low-skilled office workers earning around £300 a week. Given this information, if on the day of her job loss the newly unemployed worker sets the weekly wage she aspires to at £1,000, she will choose to be unemployed, at least to start with, because she does not wish to fill a lower-paid vacancy. The lower weekly wage on offer, and perhaps poorer conditions of work and status associated with the lower-paid job, fail to meet her aspirations. She may also realise that she possesses imperfect information about the state of the job market. This means she needs to search the labour market to find out whether better-paid and higher-status vacancies exist, which she does not know about currently.

Approached in this way, frictional unemployment can be viewed as a voluntary search period in which newly unemployed workers scan the labour market, searching for vacancies which meet their aspirations.

There are a number of ways in which a voluntary search period can end. First, in the example described above, the person may eventually learn of a vacancy for which she is qualified, and which meets her initial aspiration. Indeed, the vacancy might have been there all the time but, until she searched the job market, she was unaware of its existence. Second, the vacancy may have arisen during her search period, perhaps because of a general improvement in the condition of labour markets. Third, she may end her voluntary unemployment if she decides, on the basis of her lack of success in getting a job, that her initial aspirations were unrealistically high and that she must settle for a lower-paid, less attractive job.

Long search periods, which increase the amount of frictional unemployment in the economy, result in part from the welfare benefit system. Without the receipt of welfare benefits, search periods would have to be financed by running down stocks of saving, or through the charity of family and friends. In this situation, the threat of poverty creates an incentive to search the job market more vigorously and to reduce the aspirational wage levels of the unemployed.

The availability of a state safety net provided by unemployment benefits and other income-related welfare benefits, together in some cases with redundancy payments, enables unemployed workers to finance long voluntary search periods. For this reason, many free-market economists support a reduction in the real value of unemployment benefits, together with restricting benefits to those who can prove they are genuinely looking for work. Free-market economists believe these policies create incentives for the unemployed to reduce their aspirations quickly, which shortens search periods. Other economists would point out that this safety net ensures that the market failure caused by extreme inequality and poverty is minimised as far as possible. In other words, the welfare benefits are necessary to prevent other (some would say, bigger) problems occurring.

Structural unemployment

Structural unemployment, which is more long term than frictional unemployment, results from the structural decline of industries which are unable to compete or adapt in the face of changing demand or new products, and the emergence of more efficient competitors in other countries. Structural unemployment is also caused by changing skill requirements as industries change ways of producing their products. In the latter case, the structural unemployment is often called technological unemployment.

Technological unemployment results from the successful growth of new industries using labour-saving technology. In contrast to mechanisation (workers operating machines), which has usually increased the overall demand for labour, automation (machines operating other machines) reduces the demand for labour. Whereas the growth of mechanised industry increases employment, automation of production can lead to the shedding of labour, even when industry output is expanding.

Changes in technology can also lead to changes in the pattern of demand. In countries with hot climates, solar panels have been replacing electricity generated by coal, which leads to structural unemployment when fossil-fuel burning power stations close down.

The growth of structural unemployment in the service sector has partly been caused by the increasing use of information and communication technology (ICT) and automated services. Call centre employment has grown significantly in the service sector in recent years, though to some extent call centres have moved overseas to lower-waged countries such as India. Moreover, some companies which used to employ call-centre workers now use automated communication software rather than humans to provide customer services.

It is not always easy to separate changes in structural unemployment from other causes of unemployment, particularly changes in aggregate demand. Manufacturing output grew in many of the years between 1993 and 2008, but manufacturing employment fell. However, during these years, there was a danger of exaggerating the growth of unemployment in manufacturing industries because many activities, ranging from cleaning to information technology maintenance, previously undertaken 'in-house' by manufacturing firms, were outsourced to external service-sector providers.

KEY TERM

structural unemployment long-term unemployment occurring when some industries are declining, even though other industries may be growing. It also occurs within a growing industry if automation reduces the demand for labour, and when production requires new skills not possessed by the workers who lose their jobs. Structural unemployment is associated with the occupational and geographical immobility of labour.

STUDY TIP

Frictional and structural unemployment have supply-side causes which affect the position of the long-run aggregate supply curve. The larger these types of unemployment are, the further to the left the long-run aggregate supply curve will be positioned.

TEST YOURSELF 11.6

Explain the difference between frictional and structural unemployment.

APPLICATION OF ECONOMICS IN THE REAL WORLD 11.2

Under-employment and zero-hours employment contracts

In 2014 the UK prime minister announced: 'We have reached an important milestone, with more people in work than ever before in our history.' However, while unemployment had fallen since early 2012, under-employment had decreased to a lesser degree, to around a million people, or nearly twice the pre-recession level in 2007.

Across the UK the number of people who count as under-employed — people working part time because they cannot get a full-time job, or wanting more hours in their current job, has increased for both employees and the self-employed.

The Office for National Statistics (**www.ons.gov.uk**) cites that the number of UK workers who are under-employed has broadly varied between 4% and 8% of the workforce over the last 20 years. At the start of 2020, it was around 7%, but it rose to almost 9% at the end of that year before falling back to 7% by the end of 2021.

The problem of under-employment is exacerbated by zero-hours employment contracts. Zero-hours contracts do not guarantee a minimum number of hours of employment. The number of people on zero-hours contracts rose quickly from under 250,000 in 2011 to just under 1,000,000 in 2018, peaking at just over 1,000,000 in 2020 and falling below again in 2021. Those on zero-hours contracts are likely to be younger workers (under 25 years old). They are very slightly more likely to be female workers. Sectors with the highest proportion of workers on zero-hours contracts are accommodation and food (around 25%) and health and social care (around 20%).

About a third of workers on zero-hours contracts want more hours, primarily in their current job, compared with 10% of other people in employment.

TUC general secretary Frances O'Grady has said: 'Zero-hours contracts sum up what has gone wrong in the modern workplace. They shift almost all power from the worker and give it to their boss.' But the government's business secretary has argued that 'Zero-hours contracts are valued by many employers and individuals who want flexibility in the hours they work, such as students, people with caring responsibilities and those who want to partially retire.' The shadow business secretary responded by saying: 'Ministers have watered down every person's rights at work and zero-hours contracts have gone from being a niche concept to becoming the norm in parts of our economy. Workers are unhappy that staff can simply be sent home if there is no work to be done.'



The fast-food industry has long used zero-hours contracts for its workers

Follow-up questions

1. Why are women more likely than men to be under-employed?
2. Assess the advantages and disadvantages of zero-hours employment contracts.
3. Describe one other cause of low-paid work.
4. Should the government make zero-hours contracts illegal? Justify your answer.

EXTENSION MATERIAL

International competition and structural unemployment

The growth of international competition has been a particularly important cause of structural unemployment. During the post-Second World War era from the 1950s to the 1970s, structural unemployment in the UK was regionally concentrated in areas where nineteenth-century staple industries such as textiles and shipbuilding were suffering structural decline. This regional unemployment, caused by the decline of 'sunset industries', was more than offset by the growth of employment elsewhere in the UK in 'sunrise industries'. However, in the severe recessions of the 1980s and 1990s, and also in the 2008–09 recession, structural unemployment affected almost all regions in the UK as the **deindustrialisation** process spread across the manufacturing base. Decline of manufacturing industries as a result of international competition can also lead to structural unemployment in industries which service the manufacturing sector: for example, in private security firms.

Although manufacturing output grew in the 'boom' years before the 2008–09 recession, employment in manufacturing industries often fell. Recession and the growth of cyclical unemployment caused a further fall in manufacturing employment. However, there is a danger of exaggerating the decline in employment in manufacturing because many activities, ranging from cleaning to ICT maintenance, which were previously undertaken 'in house' by manufacturing firms, have been outsourced to external service-sector providers.

Structural unemployment has occurred within the service sector as well as in manufacturing industries. For example, increasing use of ICT, automated services and the internet has meant that total employment has fallen in the travel agency industry. Call-centre employment has grown significantly in recent years, although a decline has been forecast, partly because companies are increasingly employing automated communication software rather than human beings to provide customer service and to answer telephone and internet queries.

KEY TERMS

deindustrialisation the decline of manufacturing industries, together with coal mining.

cyclical unemployment (also known as **Keynesian unemployment** and **demand-deficient unemployment**) unemployment caused by a lack of aggregate demand in the economy, occurring when the economy goes into a recession or depression.

STUDY TIP

This is an example of applying *AD/AS* analysis to illustrate an important aspect of macroeconomics. There are many other economic problems and issues that can be analysed in a similar way.

STUDY TIP

A negative output gap is linked to cyclical unemployment.

Cyclical unemployment

Cyclical unemployment, as we have mentioned, is also known as **Keynesian unemployment**. During the Great Depression in the 1930s, Keynes — but not his opponents — believed that deficient aggregate demand was a major cause of persistent mass unemployment. Free-market economists generally agree that temporary cyclical unemployment may be caused by a lack of demand in the downswing of the economic cycle. However, Keynes went further, arguing that the economy could settle into an under-full-employment equilibrium, caused by a continuing lack of effective aggregate demand. In contrast to frictional unemployment, which is voluntary, Keynes believed that cyclical unemployment is involuntary: that is, not caused by the workers themselves. As a result, the unemployed should not be blamed for their idleness.

Figure 11.14 illustrates cyclical unemployment in the context of the economy's *AD/AS* diagram. A collapse in business and consumer confidence causes the *AD* curve to shift left from AD_1 to AD_2. Keynesian economists argue that if prices and wages are 'sticky' (or inflexible), deficient aggregate demand and cyclical unemployment will persist, with the equilibrium level of national income falling from y_N to y_2.

However, anti-Keynesian or free-market economists reject this view. They assume that markets for both goods and labour are competitive; by reducing businesses' costs of production, falling wages shift the *SRAS* curve from $SRAS_1$ to $SRAS_2$. The price level falls to P_2 and output increases from y_2, back to the normal capacity level of output, y_N. According to the free-market view, cyclical unemployment is temporary and self-correcting — provided that markets are sufficiently competitive, and prices and wages are both flexible.

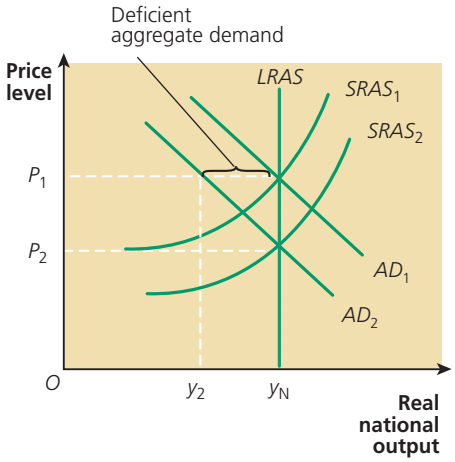


Figure 11.14 Cyclical unemployment caused by a leftward shift of aggregate demand

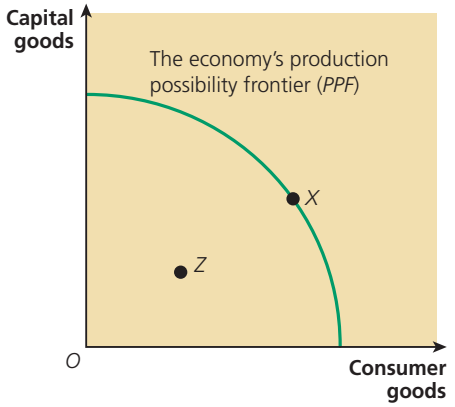


Figure 11.16 Cyclical unemployment and the economy's production possibility frontier

TEST YOURSELF 11.7

The bar graph in Figure 11.15, for a particular year, shows an estimate of the percentage of workers structurally and cyclically unemployed in the euro area (eurozone), Japan and the USA in that year. What can you conclude from the data?

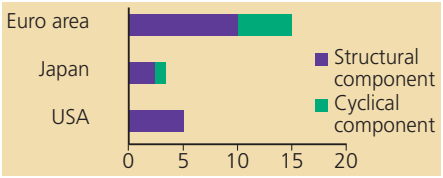


Figure 11.15 Structural and cyclical unemployment: selected areas

Figure 11.16, which shows the economy's production possibility frontier (*PPF*), illustrates another way of showing cyclical unemployment. All points on the production possibility frontier, including point *X*, show the economy using all its productive capacity, including labour. There is no demand deficiency and thus no cyclical unemployment. By contrast, deficient aggregate demand can lead to the economy producing inside its production possibility frontier: for example, at point *Z*. When this is the case, cyclical unemployment exists in the economy.

STUDY TIP

Figure 11.16 provides an example of how to use a production possibility curve to illustrate the point you are making. A *PPF* curve can also be used to illustrate scarcity, opportunity cost, choice, productive efficiency and economic growth.

EXTENSION MATERIAL

Say's Law

To explain the difference between the pre-Keynesian (free-market) and Keynesian views on the existence of deficient demand in the economy, we must introduce Say's Law, named after an early nineteenth-century French economist, Jean-Baptiste Say. In popular form, Say's Law states that supply creates its own demand. Whenever an output, or supply, is produced, factor incomes such as wages and profits are generated that are just sufficient, *if spent*, to purchase the output at the existing price level, thereby creating a demand for the output produced. Stated thus, there is nothing controversial about Say's Law; it is really a statement that is true by definition.

The controversial and critical issue concerns whether the *potential* demand or incomes generated are *actually* spent on the output produced. The pre-Keynesians believed that if households save more than firms wish to borrow to finance investment spending, the rate of interest falls to equate savings with investment. This adjustment mechanism means that Say's Law holds. Keynes disputed this, arguing that in recessionary conditions, when business and consumer confidence is low, savings will exceed investment, and there will be a glut or excess of savings. Say's Law breaks down and the resulting deficient demand causes unemployment.

KEY TERM
seasonal unemployment unemployment arising in different seasons of the year, caused by factors such as the weather.

Seasonal unemployment

Seasonal unemployment is a special case of casual unemployment, which occurs when workers are laid off on a short-term basis at certain times of the year. It occurs in trades such as tourism, agriculture, hospitality and building. When casual unemployment results from regular fluctuations in weather conditions or demand, it is called seasonal unemployment.

APPLICATION OF ECONOMICS IN THE REAL WORLD 11.3

Beveridge and full employment

In his 1944 book, *Full Employment in a Free Society*, the British economist William Beveridge said that full employment exists where there are more vacancies than there are unemployed people. This is provided the job vacancies that exist pay fair wages, which means that it would be reasonable for the unemployed to take these jobs. As a consequence, the period between someone losing one job and finding another job would be brief.

Beveridge went on to say that full employment does not mean literally no unemployment: that is to say, it does not mean that every man and woman in the country who is fit and free for work is employed productively on every day of their working lives. He argued that full employment would allow for no more than 3% unemployment due to frictions, but that there always had to be more vacancies than unemployed workers.

In terms of the categories of unemployment we have explained, Beveridge split his 3% unemployment into 1% seasonal, 1% frictional and 1% structural unemployment. In Beveridge's view, these frictions would mean that a worker was never unemployed for anything other than a very short period.

Beveridge also asked who is responsible for maintaining full employment. He believed that 'the ultimate responsibility for seeing that outlay as a whole, taking public and private outlay together, is sufficient to set up a demand for all the labour seeking employment, must be taken by the State'. Full employment required active government support to ensure that aggregate demand was sufficient to maintain employment opportunities for all those who wanted to work.

Follow-up questions

1 What is the size of UK unemployment at the time you are reading this box?
2 Which of the two definitions of full employment, the free-market definition or the Beveridge definition, has most influenced UK governments in recent years? Justify your answer.
3 Is there evidence of cyclical unemployment in the UK today? Explain your answer.
4 Why is seasonal unemployment not significant in the UK?

TEST YOURSELF 11.8
What is another name for cyclical unemployment?

Demand-side and supply-side factors affecting employment and unemployment

As we have explained, cyclical or Keynesian unemployment is caused by a lack of aggregate demand in the economy. This means that it is a demand-side form of unemployment, as for the most part is seasonal unemployment. By contrast, frictional and structural unemployment result from supply-side factors affecting the economy. Appropriate government policies to reduce the different types of unemployment depend on whether they are caused by demand-side or supply-side factors. Cyclical unemployment can be reduced through using fiscal and/or monetary policy to stimulate aggregate demand. However, supply-side policies should be used to reduce frictional and structural unemployment.

SYNOPTIC LINK

Read about the Beveridge definition and the free-market definition of full employment in Chapter 9 on pages 301–02. This will help you to understand the historic context of deciding what is meant by full employment and why it matters both in a macro sense and in terms of solving the economic problem of allocating scarce resources.

The real-wage theory of unemployment

Nearly a century ago, large-scale persistent unemployment occurred in the UK, preceding the spread of unemployment worldwide in the Great Depression of the 1930s. Much of UK unemployment in the 1920s probably resulted from the lack of competitiveness and decline of nineteenth-century staple industries such as shipbuilding and textiles. This problem was made worse by an over-valued exchange rate. However, free-market economists blamed a substantial part of the unemployment on excessively high **real wages**.

SYNOPTIC LINK

How over-valued (and under-valued) exchange rates affect the economy, and unemployment, is explained in more detail in Chapter 15.

KEY TERMS

real wages the purchasing power of the nominal (or money) wage; for example, real wages fall when inflation is higher than the rise in the nominal wage rate, and real wages rise when the nominal wage rate increases more rapidly than inflation.

real-wage unemployment unemployment caused by real wages being stuck above the equilibrium market-clearing real wage.

Unemployment caused by excessively high wage rates is called **real-wage unemployment**, or classical unemployment. This type of disequilibrium unemployment is illustrated in Figure 11.17. In the diagram, full employment (equilibrium employment) is determined where the aggregate demand for labour equals the aggregate supply of labour, shown at point *X*, at the real wage rate w_{FE}. However, suppose wages are fixed at a higher real rate, at w_1 rather than w_{FE}. At this real wage rate, employers wish to hire E_1 workers, but E_2 workers wish to supply their labour. There is excess supply of labour in the labour market, equal to the horizontal distance between points *Z* and *W*: that is, $E_2 - E_1$. (Although Figure 11.17 is a macroeconomic diagram, depicting the economy's aggregate labour market, real-wage unemployment can also be thought of as a microeconomic phenomenon, related to individual labour markets that are in decline.)

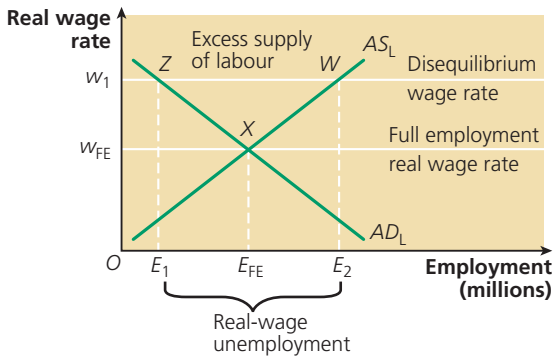


Figure 11.17 Real-wage unemployment in the economy's aggregate labour market

Free-market economists argue that, as long as labour markets remain competitive, the resulting real-wage unemployment should be temporary. Competitive forces in the labour market should cure the problem, bidding down the real wage rate to w_{FE} and thereby eliminating the excess supply of labour. Full employment should quickly be restored when the number of workers willing to work once again equals the number of people whom firms wish to hire.

But suppose labour market rigidity or 'stickiness', perhaps caused by trade unions, prevents the real wage rate falling below w_1. In this situation, the market mechanism fails to work properly, the excess supply of labour persists, and real-wage unemployment occurs. Pre-Keynesian free-market economists blamed trade unions and other causes of labour market

imperfection for the mass unemployment that occurred between the First and Second World Wars. In their view, responsibility for unemployment lay with the workers in work and their trade unions who, by refusing to accept lower wages, prevented the unemployed from pricing themselves into jobs.

Many Keynesian economists believe real-wage unemployment to be involuntary unemployment, caused by wage stickiness (where wages remain at their current level and take time to respond to changes in market conditions) or wage inflexibility over which workers have no control. By contrast, free-market economists argue that real-wage unemployment is voluntary unemployment on the grounds that workers and their trade unions should be prepared to accept wage cuts.

The natural rate of unemployment

As already noted, full employment does not necessarily mean that every single member of the working population is in work. Instead, it means a situation in which the number of people wishing to work at the going market real wage rate equals the number of workers whom employers wish to hire at this real wage rate. We also stated that even this definition needs qualifying, since in a dynamic economy there is always some frictional and structural unemployment brought about by the changes that are constantly taking place in the economy.

Frictional and structural unemployment make up what is called **equilibrium unemployment**. Equilibrium unemployment, which is also called the natural *level* of unemployment, exists even with the real wage rate at its market-clearing level. Expressed as a *rate* of unemployment, rather than as a level, this is the **natural rate of unemployment** (*NRU*).

The natural *level* of unemployment is illustrated in Figure 11.18. It is shown by the distance $E_1 - E_{FE}$ (where E_1 indicates the number of workers willing to work at the full employment real wage rate W_{FE}, and E_{FE} is the 'full employment' level of employment). The natural level of unemployment is unemployment occurring when the aggregate labour market is in equilibrium, with the aggregate demand for labour equalling the aggregate supply of labour ($AD_L = AS_L$). In Figure 11.18, point *X* shows equilibrium *employment*, with the equilibrium wage rate or market-clearing real wage rate at W_{FE}. Full employment occurs when E_{FE} workers are hired. The distance between AS_L and the curve AS_{LN} shows the amount of frictional and structural unemployment in the economy at the full employment wage rate W_{FE}, namely the number of workers who are willing and able to work at this wage rate, but who for frictional and structural reasons cannot get jobs.

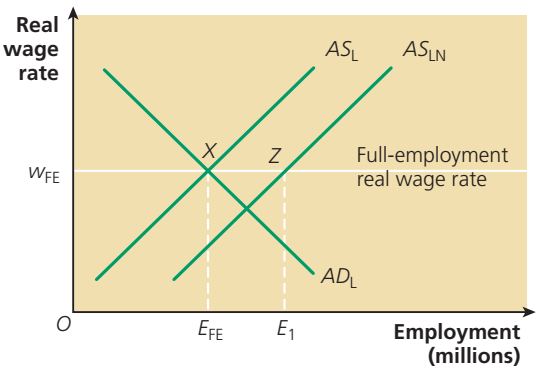


Figure 11.18 The natural *level* of unemployment, shown as $0E_1 - 0E_{FE}$, is made up of frictional and structural unemployment

TEST YOURSELF 11.9
What is another name for real-wage unemployment?

KEY TERMS
equilibrium unemployment exists when the economy's aggregate labour market is in equilibrium. It is the same as the natural *level* of unemployment.

natural rate of unemployment the rate of unemployment when the aggregate labour market is in equilibrium.

SYNOPTIC LINK
We develop the concept of the natural rate of unemployment in section 11.4 of this chapter when exploring possible conflicts between macroeconomic policy objectives.

The consequences of unemployment

The consequences of unemployment for the economy

Unemployment is bad for the economy as a whole, largely through the waste of human capital. When workers are unemployed, not all the economy's productive resources are used to produce output, which, if produced, could add to the material standards of living and economic welfare of the whole population. Instead, the economy produces inside its production possibility frontier and fails to operate to its potential.

Unemployment is also one of the factors that reduces an economy's international competitiveness. High unemployment can reduce incentives for firms to invest in new state-of-the-art technologies that generally lead to increased export competitiveness. The under-investment associated with high unemployment also results from a reduced need to invest in capital-intensive technologies when there are plenty of unemployed workers who are not only available but cheap to hire. In other words, high unemployment means the relative cost of employing labour is low due to less pressure on wages, and this means there is a reduced need to use capital instead of labour for production. In these circumstances, employers continue to use labour-intensive but outdated technologies, particularly when high unemployment accompanies a stagnant economy, low profits and a climate of business pessimism.

Under-investment can also be caused by the higher business taxes that firms may have to pay to help finance the welfare benefits paid to unemployed workers. While it is true that the Jobseeker's Allowance can be claimed only in the first months of unemployment, in the UK the state continues to pay Universal Credit to families in which there is no wage earner, in order to save family members, particularly children, from the effects of absolute poverty. In 2013, at the time of the introduction of Universal Credit, which is replacing Income Support, the government introduced a benefit cap. This cap limits the maximum amount people can claim in benefits each year. The cap for couples and single parents is currently £20,000 for those who live outside London, and £23,000 for those who live in London. For single adults the cap levels are, respectively, £13,400 and £15,410 outside and within London. The government has justified the cap on the grounds that families living solely on benefits should not receive welfare payments higher than the earnings of people on median incomes.

Economies are particularly badly affected by long-term unemployment. A worker may become increasingly unemployable the longer that worker is out of work: for example, because of the erosion of job skills and work habits. Long-term unemployment is also made worse by the fact that employers, who might otherwise hire and retrain workers who have been economically inactive for several years, perceive that workers with more recent job experience present fewer risks and are more employable. When inactive workers are seen as unemployable, the economy begins to behave as if it is on its production possibility frontier, even though there are plenty of unemployed workers notionally available for work. An increase in aggregate demand can then lead to inflation rather than to an increase in output and jobs.

Nevertheless, despite the disadvantages of high unemployment for the economy, many free-market economists believe a certain amount of unemployment is necessary to make the economy function better. In

particular, by providing downward pressure on wage rates, unemployment can reduce inflation. Unemployment also contributes to a widening of income differentials between better-paid and low-paid workers. Some free-market economists argue that this is a good thing, believing that differences in pay are needed to promote incentives, which then create the supply-side conditions in which the economy can prosper.

The consequences of unemployment for individuals

Unemployment is obviously bad for the unemployed themselves and for their families, largely because of the way in which the low incomes that accompany unemployment lead to low standards of living. However, the costs of unemployment for the unemployed go further than this. Unemployment can destroy hope in the future. Unemployed people may feel marginalised, and their self-esteem is reduced. Families suffer increased health risks, greater stress, a reduction in their quality of diet, and an increased risk of marital break-up and social exclusion caused by loss of work and income.

STUDY TIP

Don't confuse the *causes* (e.g. the types, such as cyclical) and *effects* (e.g. the costs to individuals, families and to the economy) of unemployment.

Government policies to reduce unemployment

When governments intervene to reduce unemployment, the appropriate policy depends on identifying correctly the underlying cause of unemployment. For example, if unemployment is incorrectly diagnosed in terms of demand deficiency, when the true cause is structural, a policy of fiscal or monetary expansion to stimulate aggregate demand will be ineffective and inappropriate. Indeed, reflation of demand in such circumstances would probably create excess demand, which raises the price level with no lasting beneficial effects on employment.

In the long run, the only really effective way to reduce unemployment is to achieve successful economic growth, which increases firms' demand for new employees. This, however, is easier said than done. Free-market economists believe that growth is facilitated by anti-interventionist supply-side policies and by reducing intervention in markets. Chapter 13 explains this in detail. Keynesian economists, by contrast, argue that expansionary fiscal and monetary policy must be used to reduce demand-deficient or cyclical unemployment such as occurred during and after the 2008–09 recession.

Reducing frictional and structural unemployment

Governments can try to reduce frictional unemployment by improving the geographical and occupational mobility of labour, by reducing workers' search periods between jobs, and by introducing supply-side policies.

Improving the geographical mobility of labour

Geographical mobility could be improved by making it easier for families to move house from one region to another: for example, by subsidising

removal costs. However, the widening difference in UK house prices between south and north is in fact reducing the geographical mobility of labour. Government spending on rented social housing in areas where there are ongoing labour shortages would perhaps be the most effective way of improving the geographical mobility of labour, though for cost and other reasons, this policy is unlikely to be adopted, at least on a scale sufficient to have a significant effect.

Improving the occupational mobility of labour

Governments can improve the occupational mobility of labour by providing retraining schemes and improving information about jobs, so that people can find jobs requiring their skills-set outside of their local area. Government retraining schemes are usually less effective than those run by private-sector firms. But a problem is that employers in trades such as plumbing often prefer to avoid spending on training their employees, instead poaching newly trained workers from the few employers who do invest in training their employees. As a result of this market failure, too few workers may end up being trained.

Reducing employment search periods

The introduction of the Jobseeker's Allowance (JSA) in 1996, paid to those who are unemployed, was an attempt to reduce search periods between jobs. The allowance creates an incentive for the newly unemployed to accept lower wage rates and to speed up the search for vacancies that meet their (now reduced) aspirations. There are various restrictions on who is eligible for receiving JSA. For example, those living with a partner who works more than 24 hours per week cannot claim JSA, and those with over £16,000 in savings (including a partner's savings) are also unable to claim JSA.

Supply-side policies

Supply-side policies (explained in Chapter 13), which try to improve the competitiveness and efficiency of markets, are now used to reduce frictional and structural unemployment. In the past, deindustrialisation led to large-scale structural unemployment, concentrated in regions of decline such as coal fields and areas previously dominated by heavy industry.

Many traditional manufacturing industries and the coal industry have now largely disappeared in the UK, so there is less scope for further structural decline in these activities. The supply-side improvements of the 1980s and 1990s also created conditions in which service industries grew to replace manufacturing. As a result, more workers were able to move from declining industries into growing ones. However, as the recent and current financial services crisis shows, service industries such as banking have themselves become vulnerable to structural decline and to overseas location.

STUDY TIP

You must appreciate that unemployment has a variety of causes and hence the appropriate policies to reduce unemployment depend on the cause.

TEST YOURSELF 11.10

Table 11.2 shows changes in the UK index of production and total employment between 2011 and 2021.

Table 11.2 UK index of production and employment rate, 2011–21

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Index of production (2019 = 100)	85.3	85.4	87.5	89.0	93.0	93.2	93.3	96.6	100.0	91.6	96.0
Employment rate (%)	70.3	71.0	71.6	72.9	73.7	74.4	75.0	75.6	76.2	75.4	75.2

Source: ONS

With the help of the data, identify and explain two significant points of comparison between the changes in the index of production and the employment rate over the period shown.

SECTION 11.2 SUMMARY

- Full employment exists when the number of workers whom firms wish to hire equals the number of workers wanting to work at current market wage rates.
- Frictional unemployment is transitional and 'between jobs' unemployment. It is also voluntary and short term.
- Structural unemployment is usually long term and results from the structural decline of industries and from changes in required job skills.
- Cyclical unemployment is also known as Keynesian and demand-deficient unemployment.
- Keynesian economists argue that cyclical unemployment is involuntary, which means that the people who are unemployed are willing to work at current wage rates but there are not enough jobs available.
- Policies to reduce unemployment will be effective only if the causes of unemployment are correctly diagnosed.
- The costs of unemployment fall on the whole economy and on the unemployed and their families.
- Equilibrium unemployment, which is also known as the natural rate of unemployment, comprises frictional, seasonal and structural unemployment.
- The unemployment that exists when the labour market is in equilibrium is 'voluntary' unemployment.

11.3 Inflation and deflation

A brief history of recent inflation and deflation in the UK

In Chapter 10, we defined inflation as a persistent or continuous rise in the price level, or as a continuing fall in the value of money. Deflation is the opposite, namely a persistent tendency for the price level to fall or for the value of money to rise.

Fifty or so years ago in the 1970s, accelerating and highly variable rates of inflation caused acute problems in the UK economy. As a result, in the monetarist decade of the 1980s, the control of inflation was elevated to pole position among government macroeconomic policy objectives. Eventually, after 1980 when inflation touched 20%, the rate of inflation came down, though it increased again to 10% in 1990. Whether the fall in the rate of inflation was related to 'monetarist' monetary policy is debatable. Nevertheless, from 1993 until 2007, the UK inflation rate remained within 1% above or below the 2% inflation rate target set by the government, apart from on one occasion when it nudged over the 3% upper limit.

SYNOPTIC LINK

The meaning of monetarism and how it fits with the coverage of inflation is explained later in this section (page 401).

Until 2008, control of inflation was accompanied by arguably the longest period of continuous economic growth the UK has ever experienced, at least in modern times. Many other economies around the world, including the USA, experienced similar conditions of low inflation and steady growth; this period has become known as 'The Great Moderation'. However, in 2007 things started to go wrong. Along with other economies, the UK was hit by a sudden burst of cost-push inflation, mostly imported from the rest of the world via escalating oil, gas, commodity and food prices. Together with the credit crunch that began in the USA, severe cost-push inflation undermined business and consumer confidence. Then, in 2008, the problems in financial markets caused aggregate demand to collapse and the UK economy entered the deepest recession it has experienced since the Second World War. The rate of inflation began to fall, and some economists believed that deflation would replace inflation.

Although the price level fell for a couple of months, this did not really constitute 'deflation' — defined as a persistent fall in the price level. Sharp, but probably temporary, falls in oil, gas and commodity prices were mainly responsible for the inflation rate dipping below zero, but most commentators were expecting positive inflation to return sooner rather than later. Inflation from 2012 onwards remained close to its target range of 2% with a margin of 1% each side (i.e. 1–3%). It rose slightly above this briefly in late 2017, linked to the sharp fall in the exchange rate in the year following the Brexit referendum.

However, due to a combination of factors, inflation began to rise in 2021 and during 2022 it reached levels (close to 10% at the time of writing) which had not been experienced for over 30 years. The very high inflation was linked to sharp rises in the prices of domestic heating bills, petrol and food prices — though prices of most goods and services saw unexpectedly high price increases over this period. Oil prices, which remained high in 2022, had knock-on effects not just on the price of petrol for car users, but also on the cost of transporting raw materials and finished products into and around the country. The supply chain problems caused by the Covid-19 pandemic, the war in Ukraine and possibly issues with Brexit were also considered to blame for this very high inflation rate. In 2022, there was talk of 'stagflation' for the first time since the 1970s — the unwelcome combination of high inflation and stagnant (falling or flat) economic growth. Forecasters believed that there was a rising chance of a period of negative economic growth in 2023.

In 1997, the UK government gave operational independence to the Bank of England in setting interest rates. The Bank was able to change interest rates based on its own judgement to keep the inflation rate within 1% either side of the government's 2% target in the medium term. Interest rates were set by the Monetary Policy Committee (MPC), which met monthly (but now meet eight times a year) and voted on changes to interest rates. This independence was believed to have been a significant factor in helping to keep UK inflation low. However, the cost-push factors that led to very high inflation in 2022 were seen by many as outside the Bank of England's control.

Inflation, deflation and disinflation

Until the 2008–09 recession and immediately after, the overall price level has seldom fallen in western industrialised countries since the 1930s. Since the Keynesian revolution, the phrase 'to deflate the economy' has been applied in a rather loose way to describe a reduction in aggregate demand and levels of economic activity, output and employment. A deflationary policy uses fiscal and/or monetary policy to reduce aggregate demand, in order to take excess demand out of the system. In this situation, the rate of inflation is likely to fall, but remain positive. As explained earlier, this is known as *disinflation*. By contrast, *reflation* refers to an increase in economic activity and output, and a reflationary policy stimulates aggregate demand. Often, inflation is 'reflation gone wrong', stimulating the price level rather than real output and employment.

TEST YOURSELF 11.11
Explain the difference between deflation and disinflation.

STUDY TIP
Deflation can be used to mean a fall in the price level, but it can also be used to mean a policy which seeks to reduce aggregate demand. Always check carefully which meaning is meant to apply to the term 'deflation' when it is used.

Demand-pull and cost-push inflation

There are two basic causes of inflation: excess aggregate demand in the economy, and a general rise in costs of production. The former gives rise to **demand-pull inflation** (or **demand inflation**), while the latter is called **cost-push inflation** (or **cost inflation**). As the names imply, demand-pull inflation locates the cause of inflation in the demand side of the economy, whereas cost-push inflation has supply-side causes.

KEY TERMS
demand-pull inflation (also known as **demand inflation**) a rising price level caused by an increase in aggregate demand, shown by a shift of the *AD* curve to the right.

cost-push inflation (also known as **cost inflation**) a rising price level caused by an increase in the costs of production, shown by a shift of the *SRAS* (or Keynesian *AS*) curve to the left.

STUDY TIP
Don't confuse the *causes* of inflation (e.g. demand-pull or cost-push) with the *effects* of inflation (i.e. the problems it creates).

Demand-pull inflation

As its name indicates, demand-pull inflation is caused by an increase in aggregate demand. If the economy is initially producing on the economy's *SRAS* curve, but below the normal capacity level of output (which means to the left of the *LRAS* curve), the price level has to rise to persuade firms to produce

more output to meet the extra demand. In part, this is because firms incur higher costs when they produce more goods. For firms to maximise profit, higher prices are needed to reward firms for producing more output. Once the *LRAS* curve is reached, higher prices can temporarily encourage firms to produce beyond this point, but the increase in output cannot be sustained. The quantity of goods and services produced falls back to the normal capacity level of output, though there is no fall in the price level.

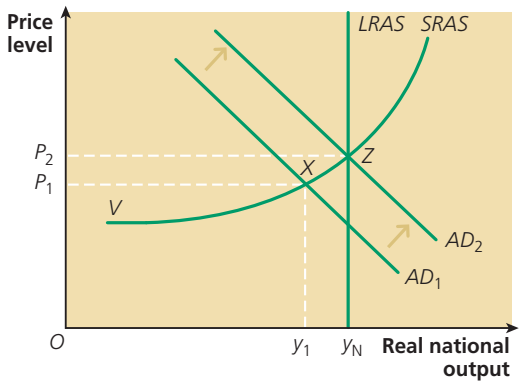


Figure 11.19 Demand-pull inflation illustrated by an *AD/AS* diagram

To remind you, the equation summarising the different elements of aggregate demand is:

$$AD = C + I + G + (X - M)$$

An increase in *any* of the components of aggregate demand, C, I, G or $(X - M)$, can lead to demand-pull inflation. Increases in consumption spending (C) by households or current government spending (G) — for example, on public-sector pay — may create the extra demand which pulls up the price level.

Investment spending and government spending on capital goods shift the *LRAS* curve rightwards. In the *long run*, a shift of the *LRAS* curve to the right can offset demand-pull inflationary pressures.

Aggregate demand/aggregate supply (*AD/AS*) diagrams can illustrate the main features of both demand-pull and cost-push inflation. Figure 11.19 illustrates demand-pull inflation.

In the graph, we have assumed that equilibrium national income is initially at point *X*. The *AD* curve is in the position AD_1; real output is at level y_1 and the price level is P_1.

Given this initial situation, any event that shifts the *AD* curve to the right — for example, to AD_2 — causes the price level to rise, in this case to P_2. In this example, real output increases to its normal capacity level of y_N. At the price level P_1, the economy's firms are only prepared to produce an output of y_1. This means that a higher price level is needed to create the conditions in which firms increase output from y_1 to y_N.

STUDY TIP

Make sure you understand why an economy may be able to produce *temporarily* above its production potential and normal capacity level of output, and to the right of the *LRAS* curve, but that this higher level of output cannot be sustained.

Figure 11.19 shows an economy initially producing below the normal capacity level of output, and then moving to normal capacity, once the *AD* curve has shifted right to AD_2. Following the increase in aggregate demand, equilibrium national income is at point *Z*, with the economy also on its long-run aggregate supply (*LRAS*) curve. In this situation, the economy is producing at normal capacity, so any further shift of aggregate demand to the right would result solely in the long run in demand-pull inflation, with no increase in real output (except possibly on a temporary basis).

Contrast this outcome with what would happen if the *AD* curve had initially been located substantially to the left of AD_1: for example, at point *V*. The economy would be in deep recession, suffering severe demand-deficient or cyclical unemployment. Given this initial position, a shift of aggregate demand to the right would increase output and employment, with relatively little effect on inflation to start with. Arguably, the adverse effect of a rising price level would be less significant than the boost to output and employment brought about by an increase in aggregate demand. However, as the increasing slope of the *SRAS* curve suggests, as the *AD* curve shifts right and moves closer to the *LRAS* curve, increasingly the *reflation* of real output (and employment) gives way to *inflation* of the price level.

Cost-push inflation

During the Keynesian era in the 1960s and 1970s, the rate of inflation increased even when there was little evidence of excess demand in the economy. This led to the development of the theory of cost-push inflation. Cost theories of inflation often locate the cause of inflation in structural and institutional conditions on the supply side of the economy, particularly in the labour market and the wage-bargaining process (known as wage-cost inflation), though rising prices of energy and/or commodities can also cause cost-push inflation (called import-cost inflation).

Cost-push theories generally argue that the growth of monopoly power in the economy's labour market and in its markets for goods and services is responsible for inflation. In labour markets, growing trade union strength in the Keynesian era enabled trade unions to bargain for money wage increases in excess of any rise in labour productivity. Monopoly firms were prepared to pay these wage increases, partly because of the costs of disrupting production, and partly because they believed that they could pass on the increasing costs as price rises when they sold output in the markets for their goods. As legislation has weakened the power of trade unions, cost-push pressure is now more likely to arise from increases in the prices of raw materials and other inputs into the production process.

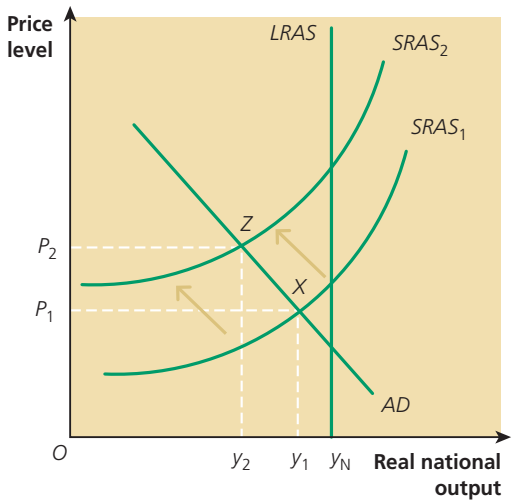


Figure 11.20 Cost-push inflation illustrated on an *AD/AS* diagram

The *AD/AS* diagram in Figure 11.20 illustrates cost-push inflation. Once again, equilibrium national income is at point *X*, with real output and the price level respectively at y_1 and P_1. In this case, we assume that the money costs of production that firms incur when they produce output rise: for example, because money wages or the price of imported raw materials increase. The increase in production costs causes the *SRAS* curve to shift left and up from $SRAS_1$ to $SRAS_2$.

As a result of the shift of the *SRAS* curve to the left, the price level increases to P_2, but higher production costs have reduced the equilibrium level of output that firms are willing to produce to y_2. The new equilibrium national income is at point *Z*. The economy could be in recession, in which case recessionary effects may moderate the cost increases.

TEST YOURSELF 11.12
With the help of an *AD/AS* diagram, explain the difference between demand-pull and cost-push inflation.

Why it is important to diagnose correctly the causes of inflation?

Just as a government must correctly diagnose the cause or causes of unemployment when implementing policies to reduce the level of unemployment, so the government (and its central bank) must first determine whether inflation is caused by excess demand or by cost-push factors when deciding on the appropriate policies to reduce or control inflation.

In the early 2000s, the Bank of England, which is responsible for counter-inflation policy, generally assumed that UK inflation is caused by excess aggregate demand. The increase in aggregate demand that occurred in the 1990s and early 2000s was accompanied by an absence of cost-push inflationary pressure. This was partly due to the success of the supply-side policies implemented in the 1980s and 1990s. These policies, which are explained in Chapter 13, improved labour market flexibility partly by attacking the power of trade unions. The UK economy was also benefiting from the benign effect of globalisation, which at the time reduced the prices of imported manufactured goods.

As long as the assumption holds that inflation is caused primarily by excess aggregate demand, raising or lowering interest rates (i.e. monetary policy) remains an appropriate policy for controlling inflation. The fact that interest rate policy kept the rate of inflation within the government's target range in virtually every month between 1997 and 2007 gave further support to the view held at the time that UK inflation had been primarily of the demand-pull kind.

More recently, the Bank of England has acknowledged that rising import prices are responsible for cost-push inflation. Indeed, cost-push inflationary pressures have become very significant, particularly those stemming from the increased prices of energy, oil and other commodities such as copper and wheat. The disruptions to supply chains caused by Covid-19, Brexit and the war in Ukraine have added to this pressure. In this situation, raising interest rates to reduce aggregate demand can be an ineffective policy for tackling cost-push inflation, unless it is argued that the economy should suffer a severe recession in order to reduce the demand for imported oil, gas and industrial raw materials. Consequently, UK governments now have to face up to the fact that they may lack the tools for controlling the rate of inflation, apart from interest rates and other monetary instruments (see Chapter 12).

EXTENSION MATERIAL

How theories of the causes of inflation have changed over the years

Table 11.3 provides a summary of how theories of inflation have developed over the years, and of how the rate of inflation has changed in the UK in recent decades.

Table 11.3 Theories of inflation and details of recent UK inflation

Eighteenth century to the 1930s	The old quantity theory of money is dominant.
1930s	The problem of inflation disappears. Keynes's *General Theory* explains deflation in terms of deficient aggregate demand.
1940 and 1950s	Keynesians develop Keynes's ideas to explain how, in conditions of full employment, excess demand can pull up the price level through demand-pull inflation.
1950s	The early monetarist theory of inflation begins to develop when Milton Friedman revives the quantity theory of money (the modern quantity theory).
1950s–1960s	Many Keynesians switch away from the demand-pull to the cost-push theory of inflation.
1960s	The Phillips curve (explained in section 11.4) is introduced into the Keynesian demand-pull versus cost-push debate.
1968	The role of expectations in the inflationary process is incorporated into the monetarist theory of inflation. The theory of adaptive expectations is built into Milton Friedman's theory of the long-run Phillips curve.
1970s	The Phillips curve relationship appears to break down with the emergence of stagflation.
1980s onwards	There is controversy once again between cost-push and demand-pull explanations of inflation, with New Keynesian explanations versus monetarist and free-market explanations, and the latter incorporating rational expectations into their models of the economy.
2008 onwards	For a short period the problem of inflation largely disappears, being replaced by the fear of deflation or a falling price level.
2015 onwards	Deflation, which had been a temporary phenomenon, gives way to a low rate of inflation which continues at least until 2019.
2021 onwards	Rapidly rising cost-push inflation in late 2021 and 2022 results from rising energy prices (gas and oil), as well as the shortages of products and increased costs connected with the disruption caused by the Covid-19 pandemic, Brexit and the war in Ukraine. High inflation means real incomes fall significantly in 2022. As a result, stagflation emerges, where prices rise and economic growth rates fall.

KEY TERMS

monetarists economists who argue that a prior increase in the money supply is the cause of inflation.

monetarism narrow monetarism centres on increases in the money supply as the prime cause of inflation; broader monetarism focuses on the virtues of free markets in resource allocation.

The quantity theory of money and monetarism

Monetarism

Around 50 years ago, a group of generally pro-free-market economists became known as **monetarists**. Monetarist economists subscribe to the demand-pull theory of inflation, but they go one stage further by arguing that excess aggregate demand for output is caused by a prior increase in the money supply. To quote the leading monetarist economist Milton Friedman, monetarists believe that 'inflation is always and everywhere a monetary phenomenon'. Monetarist economic theory is known as **monetarism**. In its narrowest sense, monetarism is the theory that inflation is caused by an increase in the money

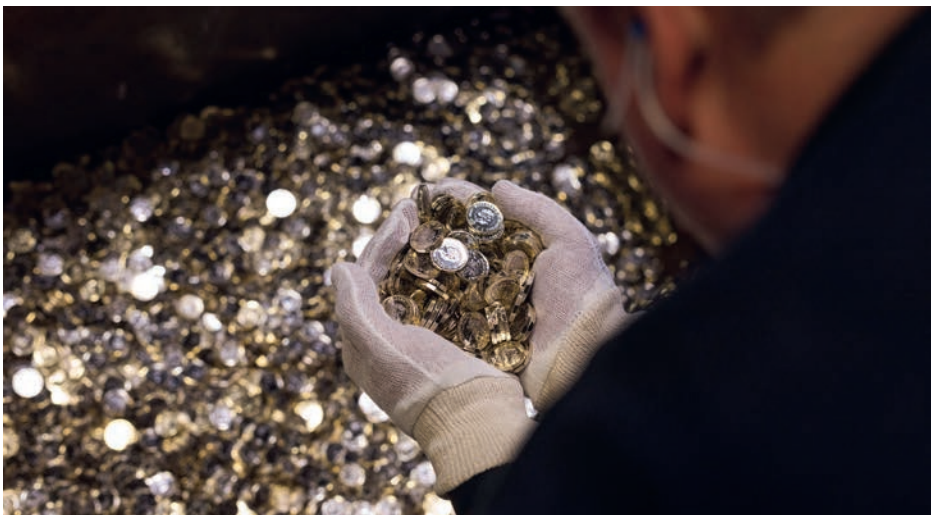
supply. In a wider meaning, it is a name given to pro-free-market economic theories and policies.

The quantity theory of money

A very old theory that originated hundreds of years ago, the **quantity theory of money**, lies at the heart of the monetarist theory of inflation. Suppose the government creates or condones an expansion of the money supply greater than the increase in real national output. As a result, this extra money, when spent, pulls up the price level — provided real output does not expand in line with the increase in spending power. At its simplest, the quantity theory is sometimes described as too much money chasing too few goods.

TEST YOURSELF 11.13
What is the difference between monetarism and a monetarist?

KEY TERM
quantity theory of money
the oldest theory of inflation, which states that inflation is caused by a persistent increase in the supply of money.



Cash in the form of notes and coins is a relatively small part of the total money supply

The equation of exchange

The starting point for developing the theory is the equation of exchange, devised by an American economist, Irving Fisher, early in the twentieth century:

$$\begin{matrix}\text{money supply}\\(\text{stock of money})\end{matrix} \times \begin{matrix}\text{the velocity of}\\\text{circulation of money}\end{matrix} = \text{price level} \times \begin{matrix}\text{quantity}\\\text{of output}\end{matrix}$$

In the equation, for a particular time period, say a year, the stock of money in the economy (M) multiplied by the velocity of circulation of money (V) equals the price level (P) multiplied by the quantity of real output (Q) in the economy. On the left-hand side of the equation, the velocity of circulation (V) is the speed at which money circulates around the economy when people use money to buy goods. Monetarists argue that V is constant or at least stable. This means that, when M increases, it is spent on goods and services. If Q is unable to increase, the price level P is pulled up by excess demand. Keynesian economists, by contrast, believe that when M increases, it may be partially absorbed by a slowdown in V, which means that much of the extra money is not spent on consumer goods and services. (However, some of the extra money might be spent on investment in new capital goods, which would have a beneficial effect on aggregate demand and economic growth, stimulating Q rather than P.) In summary, PQ, on the right-hand side of the equation, can increase, either because real output increases or because the price level increases.

STUDY TIP

Remember that the velocity of circulation of money is the speed at which money circulates round the economy when money is spent.

The equation of exchange is often written as $MV = PT$, where T stands for total transactions taking place in the economy. This is the version of the equation first stated by Irving Fisher. A transaction occurs whenever a good or service is bought and sold. Total transactions include second-hand transactions as well as the exchange of new goods and services. Because of this PQ is a better measure than PT of expenditure on national output.

EXTENSION MATERIAL

Why Keynesian economists generally reject the quantity theory of money

Keynesians generally reject the quantity theory of money as an explanation of inflation, or claim that it only provides an explanation of rising prices when a number of highly restrictive assumptions hold. There are three main ways in which Keynesians have attacked the quantity theory.

The velocity of circulation is not constant

Much of the debate between Keynesians and monetarists about the quantity theory has centred on the issue of whether the velocity of circulation of money (V) is constant. In ordinary language, V represents how often money is spent. Monetarists believe that, because money earns little or no interest, it is generally rational to spend quickly any extra money holdings, either on goods or on non-money financial assets such as shares, and not to hold idle money balances for any length of time. By contrast, Keynesians take the opposite view, arguing that under certain circumstances (particularly when share and bond prices are expected to fall), it is perfectly sensible to hold idle money balances. In this situation, to avoid capital losses they expect to suffer from falling financial asset prices, people decide to hold money instead as an idle wealth asset. They hang on to, rather than spend, their extra money holdings.

Under-full-employment equilibrium

Keynesians also attack the quantity theory by arguing that, if there is spare capacity and unemployment in the economy, an increase in the money supply may increase real income and output (y and Q) rather than the price level P. In the Keynesian view, market forces are too weak and take too long to return the economy to its normal capacity output. This means that the economy can get stuck in an under-full-employment equilibrium.

Reverse causation

The two Keynesian attacks on the quantity theory we have so far examined accept the monetarist argument that changes in MV *can* cause changes in PQ, but they argue that an increase in the money supply (M) does not *necessarily* result in inflation (an increase in P). Instead, an increase in the money supply may be absorbed in a slowing down of the velocity of circulation of money (V), or in a reflation of real income or output, which is a good thing.

The third Keynesian attack on the quantity theory is more deep-seated, since it is based on the idea of *reverse causation*. Instead of changes in the money supply causing changes in the price level, the true relationship is the opposite: changes in the price level cause the money supply to change. In this interpretation, inflation is caused by cost-push institutional factors in the real economy. The money supply then passively adapts, expanding to the level required to finance the level of desired transactions that the general public undertake at the new higher price level. In essence, the money supply accommodates itself to (rather than determines) the price level.

Keynesians agree with the monetarists over what they consider to be the rather trivial point that an increase in the money supply is needed to finance a higher price level and allow inflation to continue. However, the reverse causation argument rejects the view that an increase in the money supply is the cause of inflation. The monetarist response is that if the central bank were to exert control over the money supply, a temporary rise in the price level caused by rising costs could not persist.

Keynesians argue that if a government tightly restricts the growth of money to try to stem inflation, the main effect might be that the current level of transactions cannot be financed, so real activity will fall, resulting in higher unemployment. This effect occurred in the credit crunch that started in 2007, though the cause of the credit crunch lay not in tight control of the money supply by the government, but in a sudden collapse in the supply of money or liquidity emanating from the banking system itself.

The role of expectations in the inflationary process

TEST YOURSELF 11.14
Explain the meaning of the velocity of circulation of money.

It is now widely accepted by economists that people's expectations of *future* inflation can affect the *current* rate of inflation. Along with the quantity theory of money, expectations of future inflation are an important part of the monetarist theory of inflation. The leading monetarist Milton Friedman was one of the first economists to draw attention to the role of expectations in the inflationary process.

Theories of expectation formation are complicated. Two theories of expectations formation are described in the extension material which follows. However, the central idea is simple: if people expect that the rate of inflation next year is going to be high, they will behave in an inflationary way now, and their behaviour will deliver high inflation next year. Trade unions and workers bargain for higher wages, and their employers then raise prices, in anticipation of tomorrow's higher expected inflation rate. Workers and firms try to 'get their retaliation in first', to avoid being left behind when the inflation rate they are expecting eventually materialises.

Likewise, when people expect the inflation rate to fall, they behave in a way that enables low inflation to be achieved. Governments therefore try to talk down the rate of inflation by convincing people that government policies are credible and that the government (and its central bank) know how to reduce inflation. The UK government's decision to make the Bank of England operationally independent in 1997 was part of an attempt to convince people (and financial markets) of the *credibility* of government policies and of its determination to keep the rate of inflation low. However, when the Bank fails to hit the 2% inflation rate target, credibility is eroded. Indeed, the expectations for future inflation in the UK have recently risen, which suggests a loss in credibility for the Bank of England being able to achieve its inflation target. This may make it harder to get actual inflation back to target as people build their higher expectations of inflation into higher wage claims.

Inflation psychology

One of the factors that has made inflation difficult to control in the UK has been the existence, built up over decades, of an 'inflation psychology'. Over the years, many groups in British society, including house owners, wage earners in strong bargaining positions, and also governments, have done extremely well out of inflation. Home owners with large mortgages, and also the government, have a vested interest in allowing inflation to continue in order to reduce the real value of their accumulated debt. (Indeed, property owners do even better when house price inflation exceeds the general rate of inflation. In this situation, the real value of houses increases while the real value of mortgages falls.)

Between 1997 and 2007 UK governments successfully cut through much of this inflation psychology by convincing people that inflation would remain low and around the target inflation rate of 2%. This benign effect on people's behaviour and their expectations made it much easier to control inflation. However, even in these years, some economists argued that circumstances could change quickly for the worse, and that inflationary dangers should be regarded as dormant rather than dead.

Inflation, as we have seen, can rise again and the full effects cannot always be anticipated as shocks to the economy become more frequent. The inflation of 2022 was an unhappy coincidence of the effects of more than one shock overlapping.

EXTENSION MATERIAL

Two theories of expectations formation

The two theories about how people form their expectations of future inflation are called the *theory of adaptive expectations* and the *theory of rational expectations*.

Adaptive expectations are where people form their beliefs of future events based on past events. For example, expectations of the future inflation rate are based on the past inflation rate. As the past and current inflation rates change, expectations of future inflation will change as a result. If inflation has been high in the past, a person using adaptive expectations will continue to expect inflation to remain high.

Rational expectations are where people form their expectations using all available information that exists to support decision making. In effect, people have a complete understanding of how the economy works. They will know that if a government is pursuing expansionary policies, inflation is likely to rise in the future — even if it is currently low. Any new information will mean that people immediately update their future expectations of any economic variable.

Modern free-market economists believe that it is unrealistic to assume that a rational economic agent, acting on self-interest, forms expectations of future inflation solely on the basis of past or experienced inflation. Self-interest requires quick modification of economic behaviour in line with expectations formed on the basis of the most up-to-date information available.

The first stage in the development of the monetarist theory of inflation was Milton Friedman's revival of the quantity theory of money in the 1950s. The second stage, again developed by Milton Friedman, was the incorporation of the theory of adaptive expectations into the monetarist theory. We will explain this shortly in relation to the Phillips curve.

The incorporation of the theory of rational expectations into the explanation of the inflationary process represents the third stage in the development of monetarist and free-market theories of inflation.

The consequences of inflation for the performance of the economy and for individuals

STUDY TIP

You should discuss the benefits and costs of inflation when explaining the effects of inflation on the economy.

Inflation can impose serious costs on both the economy and individuals, and the seriousness of these costs depends on whether individuals successfully anticipate the inflation rate. If inflation could be anticipated with complete certainty, some economists argue that it would pose few problems. Households and firms would simply build the expected rate of inflation into their economic decisions, which would not be distorted by wrong guesses.

When inflation is relatively low, with little variation from year to year, it is quite easy to anticipate next year's inflation rate. Indeed, creeping inflation, which is associated with growing markets, healthy profits and a general climate of business optimism, *greases the wheels* of the economy (making it run smoothly). Viewed in this way, a low rate of inflation — and not absolute price stability or zero inflation — may be a necessary side-effect or cost of expansionary policies to reduce unemployment.

However, rather than greasing its wheels, inflation may *throw sand in the wheels* of the economy (i.e. restricting it), making it less efficient and competitive. If the 'sand-in-the-wheels' effect is stronger than the 'greasing-the-wheels' effect, the costs or disadvantages of inflation exceed the benefits or advantages.

TEST YOURSELF 11.15
Is inflation always bad?

A low but stable inflation rate may also be necessary to make labour markets function efficiently. Even if average real wage rates are rising, there will be some labour markets in which real wages must fall in order to maintain a low rate of unemployment. When prices are completely stable (i.e. when the inflation rate is zero), to cut real wage rates, nominal wage rates have to fall. To save jobs, workers may be willing to accept falling real wages caused by money wage rates rising at a slower rate than inflation. However, workers are generally much less willing to accept cuts in money wage rates. Thus, with zero inflation, the changes required in relative real wage rates, which are needed to make labour markets function efficiently, fail to take place. Labour markets function best when inflation is low but stable. By contrast, absolute price stability produces wage stickiness, which results in unnecessarily high unemployment.

Some of the disadvantages of inflation

- **Distributional effects.** Vulnerable groups in society, living on fixed incomes, lose, while those in strong bargaining positions gain. This is because those with higher incomes are more likely to be in a position where they can ensure their wages rise to keep up with inflation. Lower-paid people are less likely to have this ability. Also, with rapid inflation, real rates of interest may be negative. In this situation, lenders are really paying borrowers for the doubtful privilege of lending to them, and inflation acts as a hidden tax, redistributing income and wealth from lenders to borrowers.
- **Distortion of normal economic behaviour.** Inflation can distort consumer behaviour by causing households to bring forward purchases and hoard goods if they expect the rate of inflation to accelerate. Similarly, firms may divert funds out of productive investment in fixed investment projects into unproductive commodity hoarding and speculation (buying things solely to make a profit by selling them for a higher price later). People are affected by *inflationary noise*. This occurs when changes in relative prices (i.e. a rise or fall in the price of *one* good) is confused with a change in the *general* price level or inflation.
- **Breakdown in the functions of money.** In a severe inflation, money becomes less useful and efficient as a medium of exchange and store of value. In the most extreme form of inflation, a *hyperinflation* in which the rate of inflation accelerates to a minimum of several hundred per cent a year, less efficient barter replaces money and imposes extra costs on most transactions.
- **Reduced international competitiveness.** When inflation is higher than in competitor countries, exports increase in price, putting pressure on a fixed exchange rate. Lower growth and rising unemployment are likely to result. With a floating exchange rate (where the exchange rate is allowed to move up and down in value with no attempt to intervene to influence its value), the exchange rate falls to restore competitiveness, but rising import prices may fuel a further bout of inflation.
- **Shoe leather and menu costs.** Consumers incur 'shoe leather costs' (also called 'search costs'), spending time and effort shopping around and checking which prices have or have not risen. By contrast, menu costs are incurred by firms having to adjust price lists more often. Many years ago, the only way to compare prices would have been to physically walk to different shops making comparisons — hence the wear and tear on a person's 'shoe leather'. The use of the internet makes this less of an issue and reference is often made to 'search costs' instead.

Some of these costs of inflation have become less significant in their impact. The ability to 'shop' online has minimised the impact of shoe-leather costs.

Modern technology has also made it easier for firms to adjust prices more often. For example, vending machines can now be easily adjusted to increase the price of an item — previously the prices were mechanically set to recognise the exact change only in coins.

Other costs have increased in impact. Weakening trade union power has made it easier for firms not to give in to wage demands, which increases the likelihood of real incomes falling when inflation rises. Fiscal drag — where increases in nominal income 'drag' people into paying higher amounts of income tax even if their real income is unchanged — are more likely to occur if tax bands are not updated in line with inflation, as happened in 2021–22 and also 2022–23.

The consequences of deflation

Common sense might suggest that if inflation is generally seen to be bad, its opposite, deflation or a falling price level, must be good, both for the performance of the economy and for individuals. However, extended price deflation may bring its own problems. When people believe prices are going to fall, they postpone big-ticket consumption decisions: for example, replacing their cars. This may erode business confidence and trigger recession or deepen and lengthen an already existing recession. However, this assumes that falling prices are the result of a *bad* or *malign* deflation rather than a *good* or *benign* deflation.

The difference between the two is illustrated in Figure 11.21. A good or benign deflation, shown in Figure 11.21(a), results from improvements in the economy's supply side, which reduce business costs of production. Both the *SRAS* curve and the *LRAS* curve shift to the right and, assuming the *AD* curve does not itself shift, the price level falls, but output and employment rise. However, in the recessionary conditions existing in the UK economy in 2009, it was much more likely that a falling price level would signal a bad deflation. A bad or 'malign' deflation, shown in Figure 11.21(b), is caused by a collapse of aggregate demand, negative multiplier effects and possibly a credit crunch.

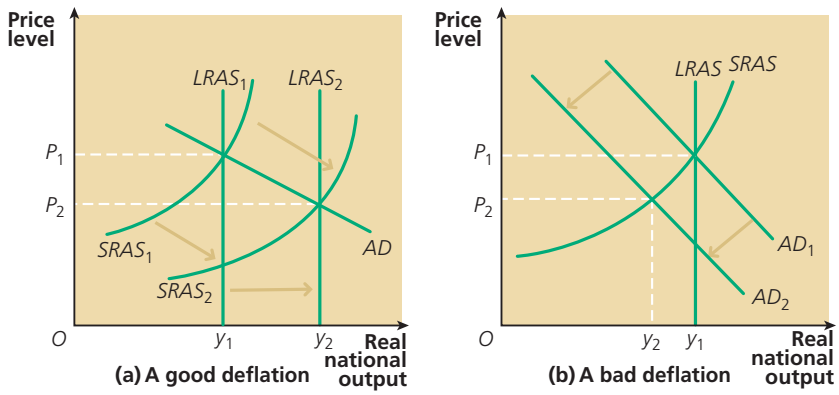


Figure 11.21 Good or benign and bad or malign deflations

STUDY TIP

Make sure you don't confuse deflation and disinflation. The former refers to falling prices whereas the latter refers to falling inflation (i.e. rising prices but at a lower rate).

APPLICATION OF ECONOMICS IN THE REAL WORLD 11.4

Are lower prices a good thing? Not really

Deflation — if it gets out of control — can be just as bad as inflation. Like inflation, deflation is a form of monetary instability. It disrupts the price mechanism, so people become confused about the true value of things. Consumers see prices are going to fall, so they defer purchases and even procrastinate on other decisions, such as getting married, in the expectation that it will be cheaper later. Pay rises are few and far between. Economic growth is almost non-existent.

During deflation interest rates may appear to be low, but may actually be high. This is explained by the difference between nominal interest rates and real ones. For example, if the Bank of England's base rate is 2% and prices are falling at 3%, real interest rates are actually 5%. The cost of borrowing is, therefore, actually higher than people think. Owing £10 is a heavier burden at the end of the year than it was at the beginning. A factory must sell more widgets, a farmer more milk and a shop more goods to meet those monthly interest payments.

When people wake up to this, they might take steps to remedy the situation quickly. For individuals, the key is the housing market. If it cracks and those with huge mortgages — secured on a dwindling asset — suddenly put their homes up for sale, a deflationary spiral could erupt. In a period of deflation, the real value of people's debt rises; this reduces net wealth and often leads to lower spending.

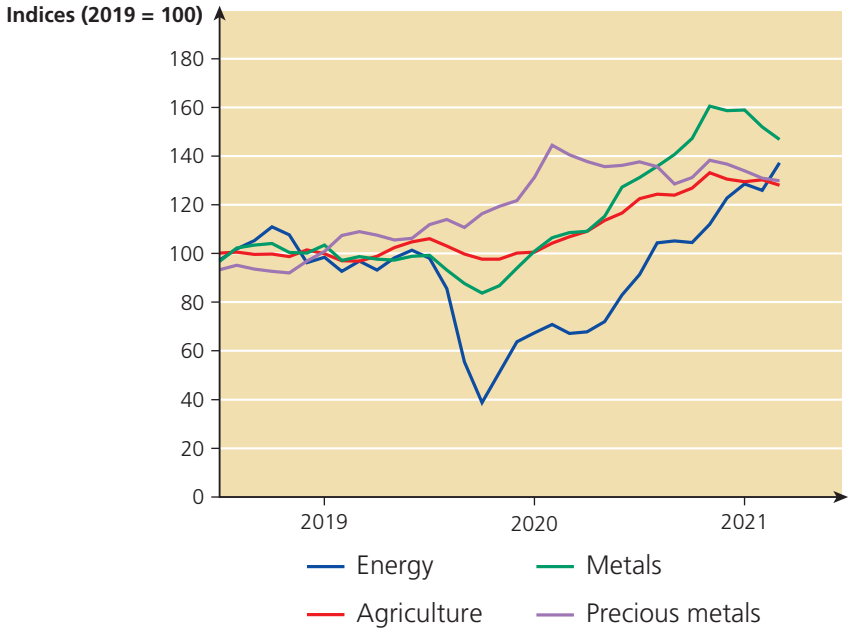
Follow-up questions

1 Who may benefit from deflation in an economy?
2 What is a 'deflationary spiral'?
3 When house prices fall, young people who have only recently bought their first houses may fall into a negative equity trap. Explain what this means.
4 In the 2008–09 recession the UK government paid motorists £2,000 if they scrapped an old car which they owned. This was nicknamed 'cash for clunkers'. What was the economic aim of this car scrappage scheme?

The effect of world commodity prices on domestic inflation

More than 40 years ago, it was generally assumed that rising wage costs were the main cause of cost-push inflation in the UK. More recently, there has been less evidence of wage-cost inflation. When it occurs, it may be largely restricted to the effect of salary increases among bankers, top business executives, premier league footballers and the like far exceeding the rise in their labour productivity. Trade unions representing public-sector workers such as train drivers on the London Underground have contributed to cost-push inflation, although from 2008 to 2017 the wages of most ordinary workers rose at a slower rate than inflation, and in some cases not at all. Pay freezes among public-sector workers since 2010 also added to the lack of wage-cost inflation.

Cost-push inflation does still exist, but it is caused mainly by the rising prices of imported food, energy and raw materials, and to a lesser extent in recent years by rising prices of manufactured goods imported from China. The resulting import-cost inflation is shown by a shift of the *SRAS* curve to the left. Figure 11.22 shows in index number form how world commodity prices changed between January 2019 and September 2021.



Source: World Bank commodity price statistics

Figure 11.22 Changes in world commodity prices, January 2019–September 2021 (2019 = 100)

TEST YOURSELF 11.16
How does a fall in world commodity prices usually cause the UK rate of inflation to fall?

There have been periods over the last ten years, such as in 2014–15 and 2020, where UK inflation was brought down by falling prices of food, oil and other commodities. However, as was the case during 2021 and 2022, rising prices in commodities will quickly feed through to rising prices. Many believe that, although gas price rises will eventually 'level off', they are unlikely ever to return to the low levels of most of the last ten years.

TEST YOURSELF 11.17
The data in Figure 11.22 are presented in the form of index numbers. Were the prices of food and all the other commodities the same in 2019? Were energy prices lower than metal prices for most of the data period? Explain your answers.

How do changes in other countries affect UK inflation?

The answer to this question is partly provided by the information above. However, other arguments can also be made.

First, the UK is now just a small part of a globalised system in which it has very little power to influence the state of the world economy. When times are good in the world economy, with rising demand for commodities and hence rising prices, the UK tends to import inflation from booming economies in the rest of the world. The prices of imported food and raw materials increase, which leads to the UK suffering import-cost inflation. Conversely, when times are bad in the world economy, with other countries experiencing recession, pressure on UK inflation is reduced. In the extreme, falling worldwide demand for UK exports, and the associated leftward shift of the *AD* curve, could lead to falling prices or deflation. Disinflation, which as we have seen is a fall in

the rate of *positive* inflation, is, however, more likely to occur than *negative* inflation or deflation. Finally, a fall in the pound's exchange rate against other currencies contributes to imported cost-push inflation.

Second, two countries, the USA and China, are particularly significant in how they can affect the UK price level. It has been said that 'when America sneezes, the rest of the world catches a cold'. The USA is so important to the UK as a source of export demand, inward investment, and business and consumer confidence, that an American downturn can exert downward pressure on the UK inflation rate. Conversely, a booming American economy, as in 2018, tends to have the opposite effect.

China's influence on the UK inflation rate is a little different. For many decades, British consumers benefited from falling prices of manufactured goods imported from China. This reduced the UK's rate of inflation. However, more recently, costs of production have risen in China, partly as a result of increased wages paid to China's labour force. Rising prices of manufactured goods, imported from China and other emerging-market economies, are increasing retail prices in the UK. As yet, however, the effect has been mild.

Outside shocks spreading from other countries also affect UK inflation. A sudden rise in oil prices caused by events in the Middle East will increase UK inflation. Conversely, the rapid growth of oil fracking in the USA has dampened world oil prices.

STUDY TIP

Make sure you are aware of important events taking place in the world economy, particularly in the European Union and in the USA. Use of recent events is one way to add the skill of application to your answers — one of the A-level assessment objectives.

SECTION 11.3 SUMMARY

- Inflation is a continuing or persistent rise in the average price level.
- In the UK, the retail prices index (RPI) and the consumer prices index (CPI) are used for measuring inflation.
- The CPI is now used for the indexation of welfare benefits, as well as for setting the inflation rate target.
- Demand-pull inflation results from the fact that, when aggregate demand increases, firms are only prepared to produce and supply more output if prices increase.
- The monetarist theory of inflation, which stems from the quantity theory of money, is a version of the demand-pull theory.
- Cost-push inflation results from higher costs of production experienced by businesses.
- Demand-pull inflation and cost-push inflation can both be illustrated on *AD/AS* diagrams.
- Deflation is a continuing or persistent tendency for the price level to fall.
- Deflation is usually thought to be bad, but there can be good deflations.
- Disinflation, which is often confused with deflation, is a slowing down in the rate of inflation.
- It is important to diagnose correctly the cause(s) of inflation when selecting policies to reduce the rate of inflation.

11.4 Conflicts between macroeconomic policy objectives

In Chapter 9 we outlined the main conflicts between macroeconomic objectives, and the resulting policy trade-offs that can face governments, particularly in the short run. Over the years, UK macroeconomic policy has been influenced and constrained by four significant short-run conflicts between policy objectives. To recap, these conflicts have been between:

- the internal policy objectives of full employment and growth and the external objective of achieving a satisfactory balance of payments
- achieving low unemployment and controlling inflation
- increasing the rate of economic growth and achieving a more equal distribution of income and wealth
- higher living standards now and higher living standards in the future.

The first of these conflicts is considered in Chapter 14, while the second appears later in this chapter where the Phillips curve is introduced (pages 413–15).

With regard to the third trade-off, by taxing the rich more and then transferring the tax revenues to the poor in the form of welfare benefits, UK governments have tried to reduce income inequalities. However, free-market economists argue that such policies reduce entrepreneurial and personal incentives in labour markets, which make the economy less competitive and the growth rate slower. In the free-market view, greater inequalities are necessary to promote the conditions in which rapid and sustainable economic growth can take place.

As regards the fourth trade-off, in the short term, the easiest way to increase living standards is to boost consumption. However, this 'live now, pay later' approach means sacrificing saving and investment, which in the long run may reduce economic growth.

When economic policy objectives are mutually exclusive, it is impossible for governments to achieve all of them at the same time — at least in the short run. The result of this is that governments often trade off between policy objectives, attempting for example to combine *relatively* low unemployment with *relatively* low inflation on the grounds that full employment and absolute price stability cannot be achieved together.

Another possibility is that, over time, governments change their ranking of policy objectives. To win votes, a government may aim for low unemployment in the months before a general election, knowing that if the election is won, policy may have to be switched to the control of inflation. In the five years between the 2010 and 2015 UK general elections, the Conservative–Liberal Democrat Coalition government started with policies of severe austerity which dampened growth and slowed down job creation. However, as the 2015 election approached, the austerity programme was replaced to a certain extent with more expansionary policies. Many commentators argued at the time that, after the 2015 election, severe austerity policies would once again be resumed, which indeed they largely were.

In 2022, the government and the Bank of England faced a difficult dilemma. With rising inflation and falling economic growth, what should happen to interest rates? Higher interest rates were likely to lead to lower economic growth and possibly push the UK into recession. Leaving interest rates at their 2021 level (or only increasing them slightly) risked the cost-push inflation of 2022 turning into a more long-term problem if inflationary expectations became entrenched and created a wage–price inflationary spiral. A former governor of the Bank of England described this situation as the Bank facing a 'narrow path' to avoid recession.

Pro-free-market economists generally believe that successful supply-side policies, combined with supply-side reform in the private sector, facilitate the production of high-quality goods and services which people, in the UK and abroad, wish to buy. The recent success of car manufacturing in the UK is quoted as evidence. According to this supply-side view, in the long run, the sustained economic growth which results resolves short-run conflicts between policy objectives. The supply-side approach to economic policy is explained in Chapter 13.

Other economists think that the supply-side view is over-stated. They argue that other countries will be attempting the same measures, meaning the gains are 'cancelled out', and that some supply-side policies, such as reducing the power of trade unions, are likely to damage labour productivity as job security is worsened.

STUDY TIP

It is useful to understand the difference, and also the links, between a policy conflict and trading off between conflicting policy objectives.

TEST YOURSELF 11.18

What is meant by a policy conflict?

How negative and positive output gaps relate to unemployment and inflationary pressures

The upper panel of Figure 11.23 shows an economy experiencing a volatile economic cycle. This means that the country's output gaps, both positive and negative, are large compared to those illustrated in the lower panel of the diagram, which depicts a smoother or more stable economic cycle.

Assuming that inflation increases and unemployment falls when there is a growing positive output gap, and that the opposite is the case with a growing negative output gap, the policy conflict between achieving these two policy objectives is greatest, the more stark the difference between the size of the positive and negative output gaps.

TEST YOURSELF 11.19

Distinguish between a negative output gap and a positive output gap.

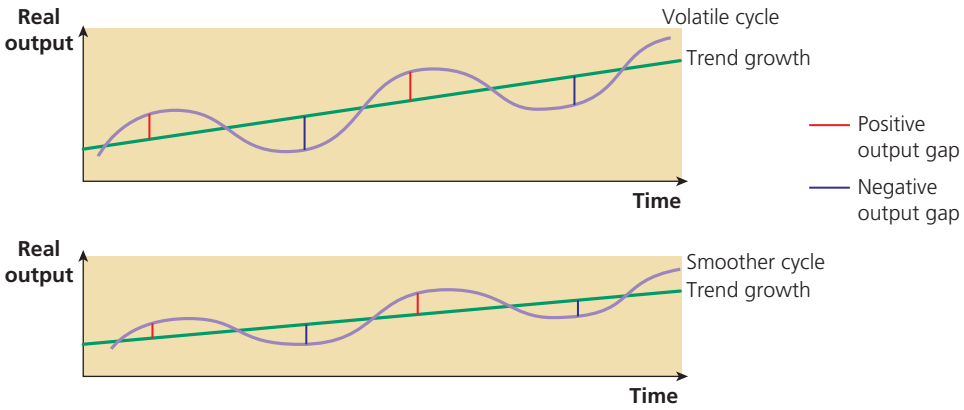


Figure 11.23 Output gaps and volatile and smooth economic cycles

Policy conflicts and the Phillips curve

The short-run Phillips curve

KEY TERM
Phillips curve based on evidence from the economy, this shows the apparent relationship between the rate of inflation and the rate of unemployment; now known as 'the short-run Phillips curve'.

The debate between economists over whether inflation is mostly caused by demand-pull or cost-push factors is sometimes conducted with the aid of a statistical relationship, the **Phillips curve**, which is illustrated in Figure 11.24.

The Phillips curve is named after the Keynesian economist A.W. Phillips, whom we mentioned in Chapter 10 on page 335 in the context of the Phillips machine. Phillips argued that statistical evidence showed that, in a period covering nearly a century, a stable inverse relationship existed between the rate of change of wages (the rate of wage inflation) and the percentage of the labour force unemployed. Later versions of the Phillips curve, such as the one illustrated in Figure 11.24, measure the inverse relationship between unemployment and the rate of price inflation.

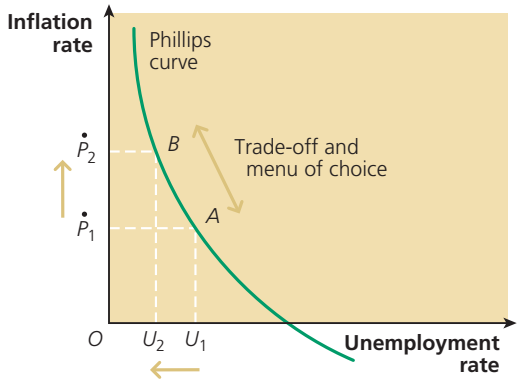


Figure 11.24 The policy choices indicated by the Phillips curve

The Phillips curve is a purported statistical relationship between two variables, and not in itself a theory of inflation. However, both the demand-pull and the cost-push theories of inflation can be used to explain the apparent relationship. In the demand-pull explanation, the factor causing unemployment to fall, moving up the Phillips curve, is excess demand, which pulls up money wages and the average price level. By contrast, in the cost-push explanation, falling unemployment means that trade union power increases, enabling unions to use their growing monopoly power over the supply of labour to push for higher wages.

Either way, the statistical evidence shown in the Phillips curve illustrates the conflict between full employment and control of inflation as policy objectives. It also suggests how the conflict can be dealt with. Suppose that initially unemployment is U_1 and the rate of inflation is $\dot{P}_1$ with the economy at point *A* on the Phillips curve. (Note the symbol $\dot{P}$ is used to show the rate of price inflation.) By increasing aggregate demand, the government can move the economy to point *B*. The unemployment rate falls to U_2, but at the cost of a higher rate of inflation at $\dot{P}_2$. By using demand management policies, it appears possible for governments to trade off between increasing the number of jobs in the economy and reducing inflation. Points such as *A* and *B* on the Phillips curve represent a menu of choice for governments when deciding an acceptable combination of unemployment and inflation.

TEST YOURSELF 11.20

Name the variables shown on the vertical and the horizontal axes of a Phillips curve diagram.

The long-run Phillips curve

Economists now generally recognise that the Phillips curve in Figure 11.24 is a short-run Phillips curve (*SRPC*), representing the short-run relationship between inflation and unemployment. In Figure 11.25, a vertical long-run Phillips curve (*LRPC*) has been added to the graph, intersecting the short-run Phillips curve where the rate of inflation is zero. The rate of unemployment at this point is the natural *rate* of unemployment (NRU), depicted by the symbol U_N. (As we explained earlier in the chapter, when expressed as the unemployment level, it is called the natural *level* of unemployment.)

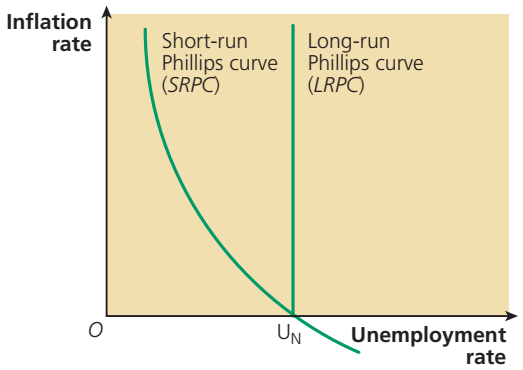


Figure 11.25 The long-run Phillips curve and the natural rate of unemployment

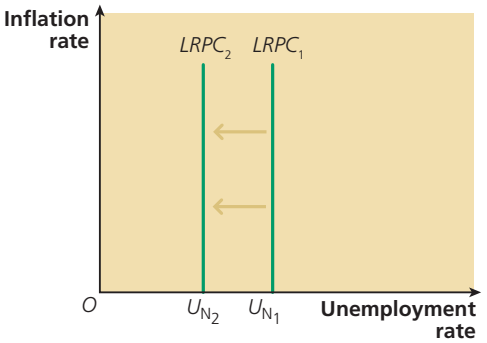


Figure 11.26 Successful supply-side policies shift the long-run Phillips curve to the left

Free-market economists argue that it is impossible to reduce unemployment *below* the natural level of unemployment, except at the cost of suffering an ever-accelerating and unanticipated rate of inflation. This would be likely to accelerate into a hyperinflation, which would then wreak severe damage on the economy. The explanation for this lies in the fact that the original Keynesian explanation of the (short-run) Phillips curve wrongly took into account only the *current* rate of inflation and ignored the important influence of the *expected* rate of inflation. If the government increases aggregate demand with the aim of reducing unemployment below the natural level, people revise upwards their expectations of *future* inflation. This then leads to higher *current* inflation. Continuous upward revision of expected inflation, of course, leads to higher and higher rates of current inflation.

The immediate solution, according to free-market economists, is for the government to respond to its mistake in trying to keep unemployment *below* its natural level, by reducing aggregate demand so as to take unemployment back to its natural level. However, the long-term solution is to reduce the natural level itself. This can be done by implementing appropriate supply-side policies (see Chapter 13), which shift the long-run Phillips curve to the left. In Figure 11.26, the curve shifts from $LRPC_1$ to $LRPC_2$, and the natural rate of unemployment falls from U_{N1} to U_{N2}. In essence, supply-side economists argue that, provided the 'correct' supply-side policies are implemented, the policy objectives of reducing unemployment and controlling inflation are compatible in the long run, though not necessarily in the short run.

EXTENSION MATERIAL

The link between short-run and long-run Phillips curves

To explain Milton Friedman's theory of the long-run Phillips curve, we must reintroduce the role of expectations into the inflationary process, and also Milton Friedman's concept of the natural rate of unemployment (NRU). As we have noted, the original Keynesian explanation of the (short-run) Phillips curve wrongly took into account only the *current* rate of inflation and ignored the important influence of the *expected* rate of inflation.

To explain the link between short-run and long-run Phillips curves, we start by assuming, as a simplification, that the rate of growth of labour productivity is zero and that the rate of increase of prices (*price inflation*) equals the rate of increase of wages (*wage inflation*). In Figure 11.27, the economy is initially at point *A*, with the unemployment rate at its natural rate U_N. At point *A*, the rate of inflation is zero, as is the rate of increase of money wages. We also assume that people form their expectations of future inflation in the next time period solely on the basis of the current rate of inflation. Thus, at point *A*, current inflation is zero, so workers expect the future rate of inflation also to be zero.

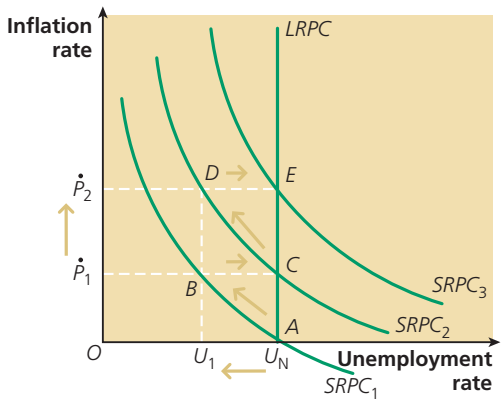


Figure 11.27 Long-run and short-run Phillips curves

We now assume that the government increases aggregate demand, to trade off along Phillips curve $SRPC_1$ to a point such as *B* in Figure 11.27, where unemployment at U_1 is below the natural rate, U_N. Inflation initially rises to $\dot{P}_1$ (say, 5%). But a point such as *B* is unsustainable. For workers to be willing to supply more labour, the real wage must rise, yet a rising real wage causes employers to demand less labour. In the short run, more workers may indeed enter the labour market in the false belief that a 5% increase in money wages is also a real wage increase. (This false belief is an example of *money illusion*, which occurs when people confuse the *real* values of economic variables with their *nominal* values.) Similarly, firms may be willing to employ more labour if they also suffer money illusion, falsely believing that rising prices mean that sales revenues are rising faster than labour costs.

This means that, to sustain an increase in employment above its natural level, workers and employers must suffer money illusion in equal but opposite directions, thereby keeping expectations of inflation, formed in the previous time period, consistently below the actual rate to which inflation has risen. However, as workers continuously adjust their expectations of future inflation to the rising actual rate and bargain for ever-higher money wages, the short-run Phillips curve shifts outwards from $SRPC_1$ to $SRPC_2$ and so on. There is indeed a separate short-run Phillips curve for each expected rate of inflation. 'Further-out' short-run Phillips curves such as $SRPC_2$ and $SRPC_3$ are associated with higher expected rates of future inflation. Conversely, the short-run Phillips curve shifts inwards when the expected rate of inflation falls.

Free-market economists argue that, in the long run, the only way to keep unemployment below its natural level is to permit the money supply to expand and finance an ever-accelerating rate of inflation. Actual inflation always has to be above the expected rate for workers and firms to be willing respectively to supply and to demand more labour. But if this happens, accelerating inflation will eventually create a hyperinflation, which, in the resulting breakdown of economic activity, is likely to increase the natural level of unemployment. Any attempt to reduce unemployment below its natural level is therefore foolhardy and irresponsible. In the short run it leads to accelerating inflation, while in the long run it perversely increases the natural level of unemployment (and the NRU) to unnecessarily high levels.

If the government realises it made a mistake when expanding the economy to point *B*, it can stabilise the rate of inflation at 5%. Workers and employers 'see through' their money illusion and realise that they have confused money quantities with real quantities. As soon as this happens, they refuse respectively to supply and to demand the labour necessary to keep unemployment below its natural level. The economy now moves to point *C*.

Short-run and long-run Phillips curves and economic policy

This section summarises points we have already made earlier in the chapter. These are:

- A downward-sloping short-run Phillips curve seems to indicate that governments can successfully trade off between the policy objectives of reducing unemployment and achieving stable prices. Looking back at Figure 11.24, expansionary policies can reduce unemployment from U_1 to U_2, provided the government accepts that the inflation rate must rise from to $\dot{P}_1$ to $\dot{P}_2$. Alternatively, if the government wishes to reduce the rate of inflation, it can do this if it is willing to accept a higher level of unemployment.
- However, with a vertical *LRAS* curve, such trade-offs are not possible. If the government expands aggregate demand in an attempt to reduce

unemployment below its natural level, it will trigger an accelerating rate of inflation which will eventually destroy the policy. Reducing unemployment below its natural level is the 'wrong' way to reduce unemployment.

- According to free-market economists, the 'correct' way to reduce unemployment is to use supply-side policies, rather than demand management policies, to shift the *LRAS* curve leftwards, thereby reducing the natural level of unemployment.

TEST YOURSELF 11.21
What are the different policy implications of short-run and long-run Phillips curves?

STUDY TIP
Make sure you understand the difference between short-run and long-run Phillips curves. This matters as there are differences in policy used to reduce unemployment or reduce inflation based on which Phillips curve you are working with.

EXTENSION MATERIAL

A further look at adaptive expectations and rational expectations

If expectations are formed rationally rather than adaptively, people do not suffer from money illusion and any attempt by a government to reduce unemployment below its natural rate fails, since 'artificially' stimulating aggregate demand leads solely to accelerating inflation. The correct way to reduce unemployment is to reduce the natural rate itself, rather than to increase demand to try to reduce unemployment below the NRU. To do this, the government should use appropriate free-market supply-side policies.

Modern free-market economists continue to accept Friedman's concept of the natural rate of unemployment. But whereas Milton Friedman believed that, in the short run at least, governments can trade off along a short-run Phillips curve and reduce unemployment below the natural rate, the theory of rational expectations rejects this possibility.

As a result, it is in workers' and employers' interests to realise instantly any mistakes made when forming expectations, and to see through any attempt by an 'irresponsible' government to reflate the economy beyond the level of output that is produced at the natural rate of unemployment. Free-market economists believe that, in this situation, attempts by government to increase aggregate demand to stimulate output and employment are anticipated fully by private economic agents. Workers and firms modify their behaviour to offset or neutralise the effects intended by the government, so the increase in aggregate demand has no effect upon real economic activity and employment. For extreme neoclassical economists, this is the case in both the short run and the long run because they believe that market forces will ensure that output and employment are always at their natural or equilibrium levels.

Another important difference separating the adaptive and rational expectations of free-market economics is the length of time unemployment must remain above its natural rate as the cost or penalty of an attempt to reduce unemployment below the natural rate. In Friedman's theory, the economy experiences a lengthy period of unemployment above its natural rate, to 'bleed' the system of inflationary expectations built up during the period of fiscal or monetary expansion.

In contrast to this 'gradualist' theory, rational expectations theory assumes that economic agents immediately reduce expectations of future inflation, provided they believe in the credibility of a tough free-market government's commitment to reducing inflation. Believing that the government means business in pursuing tight fiscal and monetary policies to control inflation, workers and firms immediately build a lower expected rate of inflation into their wage-bargaining and price-setting behaviour. Inflation falls quickly and fairly painlessly, without the need for a lengthy period of unemployment above its natural level. In effect, a firmly free-market government reduces inflation by 'talking down' inflation. However, if credibility in government policy were to disappear, its ability to control inflation would also be diminished. People would now expect higher prices, and would alter behaviour accordingly. Expectations of higher prices would become self-fulfilling.

Reconciling possible policy conflicts in the short run and the long run

In the short run, policy conflicts between reducing unemployment and achieving price stability might be reconciled in a highly depressed economy in which output is well below its normal capacity, as illustrated on a Keynesian *AS* curve (see page 362).

However, if the economy is initially producing on its vertical *LRAS* curve (or on the vertical section of a Keynesian *AS* curve) and with unemployment at its natural level, the short-run reconciliation is no longer possible. Instead, the government should try to improve supply-side conditions in the economy in the hope of increasing labour productivity and making the economy more competitive in world markets.

SECTION 11.4 SUMMARY

- When economic policy objectives are mutually exclusive, it is impossible for governments to achieve all their policy objectives at the same time, so governments often trade off between policy objectives.
- A short-run Phillips curve illustrates the conflict between controlling inflation and reducing unemployment, but is not itself a theory of inflation.
- Monetarist and free-market economists argue that the long-run Phillips curve is vertical, located at the economy's natural rate, or level, of unemployment.
- Short-run reconciliation between conflicting possible objectives is possible, provided there is significant unused capacity in the economy.
- Free-market economists believe that long-run reconciliation is not possible and that in the long run governments should aim to increase aggregate supply in the economy.

Questions

1 Explain why it is important to identify correctly the cause of unemployment.

2 Do you agree that it is more important to reduce unemployment than to reduce inflation? Justify your answer.

3 Do you agree that it is supply-side, rather than demand-side, factors which have been the more important cause of UK unemployment in recent years? Justify your answer.

4 Explain the differences between inflation, deflation, disinflation and reflation.

5 Explain why the relationship between unemployment and inflation is likely to be different in the short run from the long run.

6 Evaluate the view that the avoidance of inflation should always be the major macroeconomic objective.



12 Financial markets and monetary policy

Like all the economies of nation states in the world today, the UK's economy is a monetary economy, in which most of the goods and services produced are traded or exchanged via the intermediary of money. This chapter begins by describing the nature and functions of money in the UK economy and then explains the structure of UK financial markets and the characteristics of some of the important financial assets that are traded in these markets. The chapter goes on to explain how bank deposits, which form the largest part of the money supply in modern economies, are created by the free enterprise banking system. It then surveys the changes that have recently taken place in UK monetary policy, showing how current monetary policy has developed out of the monetary policies implemented by UK governments and the Bank of England over the last 20–35 years.

LEARNING OBJECTIVES

These are to understand:

- the characteristics and functions of money
- the financial assets which have functioned as money, both in the present day and in the past
- the structure and functions of UK financial markets: money markets, capital markets and foreign exchange markets
- the relationship between asset prices and interest rates
- the roles of different types of bank: commercial banks, investment banks and central banks
- how a commercial bank creates credit and new bank deposits
- how the Bank of England, the UK's central bank, implements monetary policy
- the regulation of the UK's financial markets.

12.1 The structure of financial markets and financial assets

Assets and liabilities

To understand fully the nature of financial markets, money and the monetary system, it is important to understand the two words: **assets** and **liabilities**. Banknotes and coins function as both financial assets and financial liabilities. A £10 note is an asset to the person owning it, since it gives the person £10 worth of spending power. However, it is a liability for the Bank of England which issued it. Many years ago, Bank of England notes were convertible into gold. The Bank had to meet this liability and, if asked by note owners, convert

KEY TERMS

asset a resource (physical or otherwise) that can be sold for money.

liability an amount owed by a business — either a current (short-term) or non-current (long-term) liability

banknotes into gold on demand. The ability to ask the Bank of England to convert banknotes into gold ended in 1931. These days and more innocuously, the Bank's liability is to replace an old and dirty note, or a note mangled in a washing machine, with a brand-new note.

From a commercial bank's point of view, a loan (or credit) granted to a customer is an interest-earning asset, which the borrower is liable to repay. But the act of creating credit or a loan simultaneously creates a bank deposit owned by the customer to whom the loan is made. The bank deposit is the customer's asset, but it is a liability for the bank itself. The bank must honour cash withdrawals and cheques or debit card payments drawn on the deposit, which transfer ownership of part of the deposit to other people. The key point to remember is that the loan-creating process increases the bank's assets (i.e. the credit it creates) and its liabilities (i.e. customers' deposits) by equal amounts. The creation of credit (an asset from a bank's point of view) simultaneously increases by the same amount customers' deposits held in the bank (a liability from the bank's point of view).

In summary, when a bank creates an interest-earning asset (the credit it extends to its customers), it is simultaneously adding to its deposit liabilities, since it must now honour customers' withdrawals from their deposits of cash or payments they make to other people. The act of creating an asset simultaneously creates a liability. (It is worth remembering that liabilities (and capital) represent a bank's 'source of funds' whereas the assets show its 'use of funds'.)

The functions of money

KEY TERM

money primarily a medium of exchange or means of payment, but also a store of value.

Money is best defined by focusing on the two principal functions it performs in the economy. Money functions as:

- **A medium of exchange or means of payment.** The economy we live in is a monetary economy in which most of the goods and services produced are traded or exchanged via the intermediary of money, rather than through barter (i.e. swapping goods). Whenever money is used to pay for goods or services, or for the purpose of settling transactions and the payment of debts, it performs this function.
- **A store of value or store of wealth.** Money is also an asset, something people own which has value. Most people store some of their wealth in the form of money in preference to holding other *financial assets*, such as stocks and shares, or *physical assets*, such as a house or car. When stored rather than spent, money's purchasing power is transferred to the future, although over time inflation may erode money's purchasing power.

TEST YOURSELF 12.1

Name two examples of an asset and two examples of a liability.

The characteristics of money

Commodity money

Figure 12.1 shows the three main forms of money that have developed since money replaced barter. (The disadvantages of barter are explained in Application of Economics in the Real World 12.1.) The earliest form of money was commodity money. Commodities that functioned as money had an intrinsic value of their own: they yielded utility and consumer services to their

owners. Beads, shells, sharks' teeth and other commodities could be used for decorative purposes while being stored as wealth. Some commodities used as money, such as cattle, could be slaughtered and eaten.

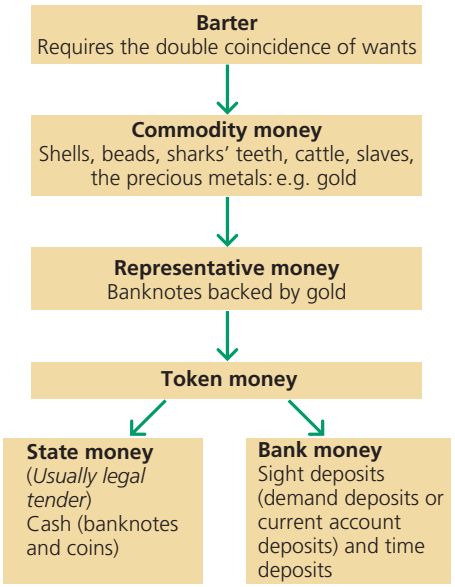


Figure 12.1 Barter and different forms of money

Representative money

As money evolved, gold and silver gradually replaced other forms of commodity money because they possessed, to a greater degree, the desirable characteristics necessary for a commodity to function as money:

- **relative scarcity** — it must be not so abundant that it is unlimited in supply (and cannot be forged)
- **uniformity** — each unit must be identical
- **durability** — it must last a reasonable amount of time
- **portability** — it must be capable of being transported by an individual
- **divisibility** — it must be capable of being divided into smaller units to facilitate transaction of a smaller value.

All of these characteristics help to create confidence in money, which is necessary for its *acceptability*. Nevertheless, gold and silver are vulnerable to theft and are difficult to store safely. Eventually, wealthy individuals deposited the precious metals they owned with goldsmiths for safekeeping.

At the next stage, the goldsmiths developed into banks, and the receipts they issued in return for deposits of gold became the first banknotes or paper money. These notes were *representative money*, representing ownership of silver or gold. Early banknotes were acceptable as a means of payment because they could be exchanged for silver or gold on demand. They were issued on the whole by privately owned banks rather than by the state, although the state continued to issue gold and silver coinage. Although worthless in themselves, banknotes functioned as money because people were willing to accept them as long as there was confidence that notes could be changed into gold, which does have an *intrinsic* value (i.e. it is seen as valuable in itself rather than just being assigned a value).

Token money

Modern money is almost all token money with no intrinsic value of its own. It takes two main forms: *cash* and *bank deposits* (money placed in a bank to be stored until a later time when the money is needed). In England and Wales, the state, or rather its agent, the Bank of England, has a monopoly over the issue of cash (although in other parts of the UK, Scottish and Northern Ireland banks also have a limited ability to issue banknotes). However, cash is best thought of as the 'small change' of the monetary system — bank deposits form by far the largest part of modern money. Most modern money takes the form of bank deposits created by the private enterprise banking system.

TEST YOURSELF 12.2
Why is most modern money 'token money'?

Money's other functions

Money has two other functions which are less important for you to know. Money serves as:

- a measure of value
- a standard of deferred payment.

Combined, these are often called the 'unit of account' function. Money is the unit in which the prices of goods are quoted and in which accounts are kept. This allows us to compare the relative values of goods even when we have no intention of actually spending money and buying goods: for example, when we 'window-shop'. This is the 'measure of value' function. Money's function as a standard of deferred payment allows people to delay paying for goods or settling a debt, even though goods or services are being provided immediately. Money acts as a standard of deferred payment whenever firms sell goods on credit or draw up contracts specifying a money payment due at a later date.

TEST YOURSELF 12.3
Explain why C is the correct answer to the following question.

A car is advertised for sale at a price of £25,000. This is an example of money's function as:

A a store of wealth

B a medium of exchange

C a measure of value

D a standard of deferred payment

EXTENSION MATERIAL

What is money?

For most people, money is so desirable and so central to everyday life that what actually constitutes it hardly merits a second thought. For people living in England and Wales, money comprises coins and Bank of England notes, and any funds on deposit in banks such as HSBC and Barclays. Residents of Scotland and Northern Ireland would also include notes issued by local Scottish and Northern Irish banks. Building society deposits as well as bank deposits are now also regarded as money, although this has not always been the case. Building societies are financial institutions that fulfil many of the functions of a commercial bank but are often owned by the members rather than shareholders and often focus more on providing savers a better return, and lend mainly for those wishing to buy houses.

Where do we draw the line as to what is money? Is a credit card money? What about foreign currency such as the US dollar or the Indian rupee, given the fact that we may not be able to spend a foreign banknote or coin in the UK? Do we include financial assets (resources which can be converted into money) such as National Savings Securities, which possess some, but not all, of the characteristics of money?

Consider also the social relationship that takes place whenever modern banknotes are spent on goods or services. Why, for example, are shopkeepers prepared to hand over new and valuable goods to strangers, in exchange for grubby and unhygienic pieces of paper with no apparent intrinsic value of their own? The answer lies in a single word: 'confidence'. In a modern economy, people are prepared to accept such tokens in settlement of a contract or debt because they are confident that these notes and coins will also be accepted when they decide to spend them.

APPLICATION OF ECONOMICS IN THE REAL WORLD 12.1

The development of modern money

Before the development of money, exchange and trade took place in village economies, where the exchange of goods and services was based on *barter* — the swapping of goods and services. However, barter is inefficient and impractical in a more complex economic system characterised by specialisation and trade. Successful barter requires a *double coincidence of wants*, which means that a person wishing to trade a television set for a refrigerator must not only establish contact with someone with equal but opposite wants (i.e. an individual possessing a refrigerator who wishes to exchange it for a television set); they must also agree that the television set and refrigerator are of equal value (or how many TV sets equal one fridge — or vice versa).

Barter is inefficient because the time and energy spent searching the market to establish the double coincidence of wants results in unnecessary search costs (so-called shoe-leather costs) and transaction costs. If trading products is inefficient, it inhibits the development of specialisation, division of labour and large-scale production. All of these contribute to the productive efficiency and range of consumer choice available in modern and sophisticated monetary economies. Bartering still takes place in some parts of the world.

Follow-up questions

1 Why does money provide a more efficient means of payment than barter?
2 What are the advantages and disadvantages of using cash as the main form of personal wealth?
3 Explain why the use of money as a medium of exchange enables an economy to benefit from specialisation and the division of labour.
4 Explain why precious metals, such as gold and silver, have often been used in economies as money.

SYNOPTIC LINK
At this stage, it might be a good idea to go ahead to section 12.2, pages 435–37, which explains how, through their ability to create credit, private enterprise banks such as Barclays and Lloyds create bank deposits, which make up the majority of modern money.

KEY TERM
money supply the stock of financial assets which function as money.

SYNOPTIC LINK
The terms 'monetarism' and 'monetarists' were explained in Chapter 11. Monetarism was a theory that placed control of the money supply right at the centre of all macroeconomic policy.

TEST YOURSELF 12.4
Explain why A provides the correct answer to the following question.

Which one of the following increases the need to hold money on one's person?

A Inflation
B Increased use of credit cards
C Increased use of automatic transfers between deposits
D Increased use of ATM machines

The money supply

Fifty or so years ago, before the advent of *monetarism*, few economists gave much attention to the precise definition of the **money supply** or stock of money in the economy. This reflected the Keynesian view that money did not matter in the macroeconomic management of the economy. However, when monetarism replaced Keynesianism as the prevailing orthodoxy in the 1970s, money did begin to matter — particularly in the years before monetarism itself drifted out of favour around 1985. For a few years from the mid-1970s to the mid-1980s, during the 'monetarist era', control of the money supply became an important part of monetarist economic management in general and monetary policy in particular. During this period, monetarist economists devoted considerable attention to the problem of deciding which assets to include and exclude when defining the money supply.

Narrow money and broad money

Over the years, the Bank of England has used more than one definition of the money supply. These divide into measures of **narrow money** and **broad money**. Narrow money, which restricts the measure of money to cash and bank and building society sight deposits, reflects the medium of exchange function of money, namely money functioning as a means of payment. Broad money also includes other financial assets which, although stores of value, are too illiquid, at least for the time being, to function as media of exchange.

In the 1980s, different measures of the money supply were closely watched as macroeconomic policy was guided by the economic theory of monetarism. There were many different measures of the money supply, all having 'code names'. For example, M0 and M1 were narrow money measures, while M3 and M4 focused on the broad money supply.

KEY TERMS

narrow money the part of the stock of money (or money supply) made up of cash and liquid bank and building society deposits.

broad money the part of the stock of money (or money supply) made up of cash, other liquid assets such as bank and building society deposits, but also some illiquid assets. The measure of broad money used by the Bank of England is called M4.

liquidity measures the ease with which an asset can be converted into cash without loss of value. Cash is the most liquid of all assets.

STUDY TIP

It is important to understand the meaning of the word 'liquidity' in the context of money, financial markets and monetary policy. **Liquidity** measures the ease with which an asset can be converted into cash without loss of value; cash is itself the most liquid of all assets. Assets that generally can only be sold after a long, exhaustive search for a buyer, such as houses and cars, are known as illiquid. Liquidity is also affected by whether or not the conversion can take place at a pre-known value. A share issued by a public limited company (PLC) is highly liquid in that it can quickly be sold on the stock exchange, but less liquid in the sense that the seller does not know the price the asset will fetch in advance of the sale, as the value of a share can rise or fall depending on how desirable that share is.

TEST YOURSELF 12.5

Is narrow money a good indicator of the store of value function of money?

TEST YOURSELF 12.6

Explain why statement C provides the correct answer to the following question, and why statements A, B and D are wrong.

Which of the following statements is true about a highly liquid asset?

A It generally has a very limited market for its resale.

B It cannot be sold on a market.

C It can be converted into a means of payment easily without loss of value.

D It has high transaction costs associated with its sale.

EXTENSION MATERIAL

Goodhart's Law

A significant problem that contributed to the decline of monetarism as a major force influencing UK governments stemmed from what has become known as Goodhart's Law. This 'law' is named after Charles Goodhart, formerly a professor at the London School of Economics and a member (between 1997 and 2000) of the Bank of England's Monetary Policy Committee (MPC). Goodhart argued that, as soon as a government tries to control the growth of a particular measure of the money supply, any previously stable relationship between the targeted measure of money and the economy breaks down. The more successful the Bank of England appears to be in controlling the rate of growth of the financial assets defined as the money supply, the more likely it is that other financial assets,

regarded previously as *near money* outside the existing definition and system of control, will take on the function of a medium of exchange and become money.

In this way, attempting to control the money supply is rather like a person trying to catch their own shadow. As soon as they move, their shadow also moves. Although what is defined as money may be controlled, this becomes irrelevant when other financial assets become money. The difficulties of, first, defining the money supply and, second, exerting control over its rate of growth contributed to the downfall of monetarism after 1985.

KEY TERMS

shares undated financial assets, sold initially by a company to raise financial capital. Shares sold by public limited companies (PLCs) are marketable on a stock exchange, but shares sold by private limited companies (Ltds) are not marketable. Unlike a loan, a share signifies that the holder owns part of the enterprise.

bonds financial securities sold by companies (corporate bonds) or by governments (government bonds) which are a form of long-term borrowing. Bonds usually have a maturity date on which they are redeemed, with the borrower usually making a fixed interest payment each year until the bond matures (see page 428).

equity the assets that people own.

debt people's financial liabilities, or money that they owe.

The difference between equity and debt

Many people living in the UK today have some possessions which they own. These may include furniture, a car or even a house; or money, **shares** and **bonds** that they possess. Even their children may own equity: for example, toys and a savings account given to them by grandparents. **Equity** is all the assets, physical and financial, which people own. **Debt**, by contrast, is what people owe: that is, their liabilities. For many adults, this is the mortgage on their house, which is money owed to a bank or building society, and credit card debt.



Even a child may possess equity

APPLICATION OF ECONOMICS IN THE REAL WORLD 12.2

Positive and negative equity

The word 'equity' has two very different economic meanings. On the one hand, in the context of topics such as fiscal policy, poverty and the distributions of income and wealth, equity means fairness or justness. However, in the context of this chapter, equity means wealth. Consider the following situation.

A person may have taken out a mortgage (money borrowed from a bank to buy a house). This then becomes their liability. (Assets are everything you *own* that possesses value, while liabilities are everything that you *owe* to other people.)

In much of recent history, the value of mortgages has remained unchanged (or even fallen when part of the mortgages are paid off each year), but house prices have risen, at least in most parts of the UK. This means that owner-occupiers of houses have benefited from *positive equity* (what they own is greater than what they owe.)

However, in the financial crisis which hit the UK in 2007, house prices fell rapidly. This created the problem of *negative equity*, which occurs when what is owed exceeds the value of what is owned: that is, the amount owed on the mortgage is more than the value of the house. People in the negative equity trap often have crippling debt from which it is very difficult to escape.

Follow-up questions

1 Name two other assets, besides the houses they own, which are usually part of householders' equity.
2 Distinguish between positive and negative equity.
3 What is the negative equity trap?
4 In the years after 2007, young house buyers who had bought their first house in the years just before the financial crisis were the main victims of the negative equity problem. Explain why this was so.

TEST YOURSELF 12.7
How can a recession lead to a negative equity trap?

EXTENSION MATERIAL

Portfolio balance decisions

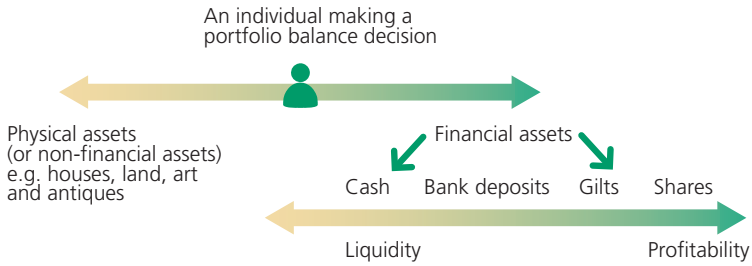


Figure 12.2 A spectrum of assets over which a portfolio balance decision may be made

Traditionally, a portfolio is a bag or piece of luggage in which people store things and transport them from one place to another. An art portfolio is the collection of drawings and paintings a student submits in order to gain an Art A-level pass. In financial language, a wealth portfolio contains the different wealth assets that an individual owns and holds at a particular point in time.

People make decisions about the form of assets in which to keep their wealth. In the first instance, people choose between holding physical assets (or non-financial assets), such as houses, and holding financial assets, such as shares. Some physical assets such as property and collectable items (e.g. rare vinyl records or antiques) can be attractive because they tend to go up in value or appreciate, thus providing some security against inflation.

When choosing the assets to hold, people make *portfolio balance decisions*. A portfolio balance decision is illustrated in Figure 12.2, which, besides making the distinction between physical and financial assets, arranges financial assets according to *liquidity* and *profitability*. As we have noted, liquidity measures the ease with which an asset can be converted into cash, and the certainty of what it will be worth when converted. We have seen that, provided it is generally accepted and can be used as a means of payment, cash is the most liquid of all assets. For the most part, bank deposits are not quite as liquid as cash, but they are sufficiently liquid to be treated as money. The other financial assets shown in Figure 12.2 are examples of non-money financial assets, in some cases being near-money. Shares and government bonds (gilt-edged securities, or gilts) are marketable (they can be sold second hand on the stock exchange), but they are less liquid than money. In contrast to money, which earns little or no interest, shares and gilts generally hold out the prospect of providing a profit for their owners.

Financial markets

In Chapter 3, section 3.1, we defined a market as a voluntary meeting of buyers and sellers. We went on to stress that markets do not have to exist in a particular geographical location, and in recent years, 24/7 global markets have grown up, facilitated by the development of the internet. This has enabled many markets to become truly global and to function on a worldwide basis.

This is certainly the case for some of the **financial markets** functioning in the UK: for example, the capital and foreign exchange markets and Lloyds insurance market. However, at the other extreme, some of the smaller financial markets, such as the market for **Treasury bills** and **commercial bills**, are restricted largely to trading in the City of London, the area of London containing many of the financial markets.

Some of the main financial markets in the UK and global economies

As explained above, a financial market is a market in financial assets or securities. (A security secures a claim against a person or institution: for example, a share secures ownership of a fraction of the company which initially sold the share to the general public.) Financial markets, which are illustrated in Figure 12.3, can be classified into:

- **Money markets** — for trading short-dated financial assets (such as Treasury and commercial bills). Assets that are bought and sold on money markets are short term, with maturities ranging from a day to a year, and are normally easily convertible into cash.
- **Capital markets** — for trading long-dated debts (such as bonds) and shares.
- **Foreign exchange markets** — for financial transactions requiring the use of foreign currency. The main participants in this market are large international commercial banks. Collectively, foreign exchange markets are the largest markets in the global economy.

KEY TERMS

financial markets markets in which financial assets or securities are traded.

Treasury bills short-dated government debt (usually three months to maturity), which pays the holder a fixed rate of interest until redemption.

commercial bills short-dated debt (usually three months to maturity), issued by private businesses, which pays the holder a fixed rate of interest.

money markets provide a means for lenders and borrowers to satisfy their short-term financial needs. The term 'money market' covers several markets, including those for Treasury bills and commercial bills.

capital markets where securities such as shares and bonds are issued to raise medium- to long-term financing, and are traded on the 'second-hand' part of the market (e.g. the London Stock Exchange and the New York Stock Exchange).

foreign exchange markets (also known as forex, FX or currency markets) global, decentralised markets for the trading of currencies.

TEST YOURSELF 12.8

What is the main difference between a financial market and a market for a good, such as the car market?

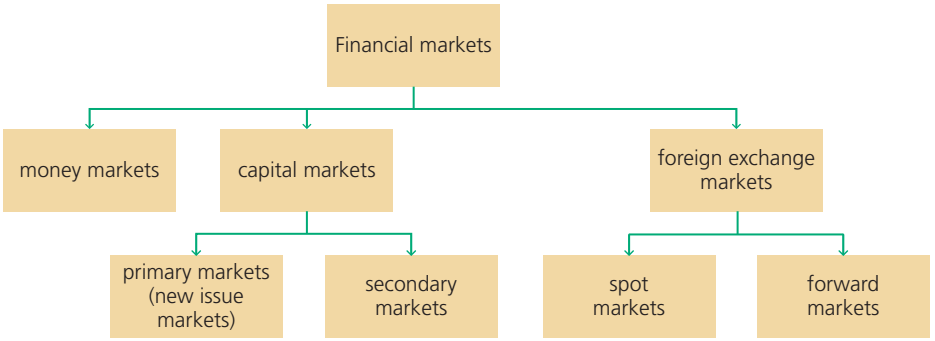


Figure 12.3 Some of the main financial markets in the UK and global economies

STUDY TIP

You should be aware of other financial markets besides those shown in Figure 12.3. These include markets for commodity futures and insurance products.

UK money markets

As we explain shortly, money markets provide mechanisms for banks to arrange their assets in terms of their liquidity or profitability, with highly liquid assets such as Treasury bills and commercial bills at the liquid end of the spectrum. Organising their assets in this way enables commercial banks to perform perhaps the most important banking function, namely that of a financial intermediary linking savers to borrowers.

A money market which has recently become extremely important in the operation of monetary policy is the London interbank market, in which the LIBOR interest rate is charged. (LIBOR is the London Interbank Offered Rate, which is a rate of interest charged when banks lend to each other, usually for very short periods of time.)

KEY TERMS

corporate bonds debt security issued by a company and sold as new issues to people who lend long term to the company. They can usually be resold second hand on a stock exchange.

government bonds debt security, in the UK known as gilt-edged securities or gilts, issued by a government and sold as new issues to people who lend long term to the government. They can be resold second hand on a stock exchange.

UK capital markets

In contrast to money markets, capital markets, which include the London Stock Exchange, provide the mechanism through which public limited companies (PLCs), which are usually the largest forms of business organisations in the UK, can raise the funds to finance their long-term growth. In addition, the part of the capital market known as the bond market performs a critical role in government finance: for example, by enabling a government to finance a budget deficit.

EXTENSION MATERIAL

Comparing money markets and capital markets

Table 12.1 enables us to compare the important similarities of, and also the important differences between, two of the most important money markets (the markets for commercial bills and Treasury bills) and the capital markets.

Table 12.1 Money markets and capital markets

	Money markets (markets for short-dated financial assets or securities)	Capital markets (markets for undated and long-dated financial assets or securities)
The private sector raising finance	Sale of commercial bills	Sale of new issues of shares Sale of new issues of corporate bonds
Central government raising finance	Sale of Treasury bills	Sale of new issues of government bonds (in the UK also known as gilt-edged securities or gilts)

The function of money markets is to supply both private-sector commercial firms and the government with a source of short-term finance. For a firm, bill finance is an alternative to a conventional bank loan. Bills are short-dated financial assets or securities, which mature within a year (and often within three months) of their date of issue. There are two main types of bill, or bills of exchange as they are known: *commercial bills* and *Treasury bills*. Commercial bills are sold by investment banks on behalf of client firms. Treasury bills, which are sold as new issues by the Bank of England on behalf of the government, provide the government with a method of financing the differences that emerge at certain times in the financial year between tax revenues and government spending.

Treasury bills are in fact short-dated government loans. (The word *bill* is used for a short-dated loan raised on a money market, whereas the word *bond* is used for a long-dated loan raised on a capital market.) The role of capital markets will be explained later in this section. Shares, corporate bonds and government bonds (gilt-edged securities or gilts) are the main securities traded on capital markets.

STUDY TIP

Don't confuse a *public* limited company (PLC) with the *public* sector of the economy. Public companies are in the *private* sector of the economy. The public sector is controlled and funded by governments.

TEST YOURSELF 12.9

How is a bond different from a share?

The inverse relationship between bond prices and interest rates

Economics students are often confused by the relationship between the prices of bonds and long-run interest rates (e.g. on government bonds). To understand the relationship, consider Figure 12.4, which shows the main characteristics of government bonds (in the UK, gilt-edged securities, or gilts) and the long-run rate of interest (such as the one paid on government debt). One of the important points to remember is that there is an inverse relationship between interest rates and bond prices — which means that, as interest rates rise, bond prices fall (and vice versa).

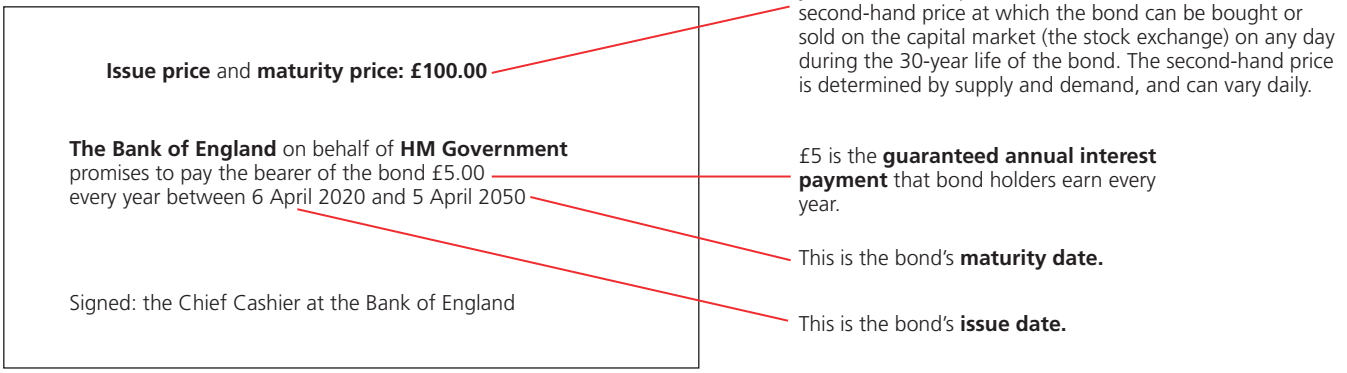


Figure 12.4 The key characteristics of a bond (in this case, a UK government bond or gilt)

KEY TERMS

coupon the guaranteed fixed annual interest payment, often divided into two six-month payments, paid by the issuer of a bond to the owner of the bond.

yield the annual interest on a bond expressed as a percentage of the bond's current market price.

maturity date the date on which the issuer of a dated security, such as a gilt-edged security (long-dated) or a Treasury bill (short-dated), pays the face value of the security to the security's owner.

Gilts (and also corporate bonds issued by private-sector companies) are *fixed-interest securities* sold by the government (or by companies) when they borrow in the long term. The bond in Figure 12.4 is a 30-year bond with a face value of £100 carrying a guaranteed annual interest payment, known as the **coupon**, of £5. Simple arithmetic will tell you that £5 is 5% of £100.

Suppose, however, that after the bond has been sold on its day of issue for £100, its second-hand price on the London Stock Exchange or bond market *rises* to £200. This means that the bond's **yield** falls to 2.5% (£5 as a percentage of £200). (A bond's yield is in effect *the long-run rate of interest* earned by the holder of the bond. It is determined primarily by the coupon and the price that the holder paid for the bond.)

The guaranteed interest (or coupon) earned each year by gilt holders is paid in two instalments at six-month intervals. Gilts and other bonds usually have a specific **maturity date**. In the case of the 4% Treasury Gilt 2016, the principal (i.e. its original face value) was repaid to gilt holders on 7 September 2016.

Suppose, by contrast, the bond's price falls to £50. This means that the bond's yield rises to 10% (£5 as a percentage of £50). Both these simple calculations illustrate the inverse relationship between bond prices and *long-run* interest rates. (However, if the calculation were made as the bond approached maturity, it would illustrate the inverse relationship between bond prices and short-run

interest rates.) If bond prices rise, yields or long-term interest rates fall, and if bond prices fall, yields or long-term interest rates rise.

QUANTITATIVE SKILLS 12.1

Worked example: calculating the yield on a bond

The government sells a new issue of 50-year gilt-edged securities, each with a nominal value, or maturity value, of £100 and a guaranteed yearly interest payment (or coupon) of £2.50. Five years after the gilt was issued, its second-hand or market price on the stock exchange falls to £50.

Calculate the yield on the gilt-edged security at this date.

The calculation is as follows:

$$\text{yield} = \frac{\text{annual coupon payment}}{\text{gilt's current market price}} \times 100$$

$$= \left(\frac{£2.50}{£50}\right) \times 100$$

$$= \left(\frac{1}{20}\right) \times 100$$

$$= 5\%$$

KEY TERM

capital gain the profit made on the buying and selling of financial and other assets (capital losses are the losses made in buying and selling assets).

EXTENSION MATERIAL

Capital gains and losses and bond and share prices

In Quantitative Skills 12.1, the fact that the gilt does not mature until 50 years after the date of issue means that an approaching maturity date has no noticeable effect on the gilt's current market price or on its yield. If, by contrast, the maturity date were next week, the market price would be very close to its maturity value (£50), and its yield would have converged to be very close to 2.5%.

For a gilt or a corporate bond with many years to go until maturity, another factor affecting changes in its market price is the expectation of making a **capital gain** (and the fear of suffering a capital loss). Suppose the current bond price is £100, but speculators *expect* the price to rise shortly to £150. If the bond can be bought now at or near its current price of £100, and its price does indeed shortly rise, speculators who buy the bond make a capital gain when selling the bond later: for example, for £150. If a sufficiently large number of speculators decide to buy the bond at or near the price of £100, increased demand for the bond pulls up the market price, until eventually it may indeed reach £150.

Conversely, a fear that the bond price will fall, say to £50, will induce speculative selling of the bond, which in turn causes the bond's price to fall. Provided a sufficiently large number of speculators behave in the same way, speculative buying and selling, in the hope of making a capital gain — or avoiding a capital loss — is an important determinant of short-run changes in bond prices. (Share prices are also affected in a similar way.)

Until 2015 there were government bonds in existence which would never mature. These were Consolidated Stock, known as 'Consols'. These bonds were attractive because of the relatively generous interest rates offered. They could be sold if investors wished to sell them. Since there are no complications brought about by an approaching maturity date, the yield on Consolidated Stock, at any point in time, is the best indicator there is of the long-run interest rate on relatively risk-free assets at that time. Long-term interest rates on risky assets such as corporate bonds and on even riskier assets, such as Greek government bonds, will be higher. As of 2014, UK Consols still in existence had been issued by Winston Churchill as chancellor of the exchequer to re-finance the bonds issued to finance the First World War. In 2014 the UK government announced that it would legally enforce redemption of all Consols, and in 2015 the last of the Consols were redeemed.

TEST YOURSELF 12.10
Explain how a capital gain differs from a capital loss.

QUANTITATIVE SKILLS 12.2

Worked example: calculating a bond's current market price

The annual coupon payment on a 40-year bond issued last year is £8. When the bond was first sold, the long-run interest rate was 8%. The bond's maturity value is £100. Within the last year, long-run interest rates have fallen to 4%.

Calculate the approximate current market price of this bond on the stock exchange.

The calculation is as follows:

$$\text{yield} = \frac{\text{annual coupon payment}}{\text{bond's current market price}} \times 100$$

$$4 = \frac{£8}{\text{bond's current market price}} \times 100$$

$$\text{bond's current market price} = \frac{£8}{4} \times 100 = £200$$

TEST YOURSELF 12.11
A £100 bond paying a nominal rate of interest of 10% with 30 years to maturity has just paid this year's annual coupon. Long-term interest rates are currently 6%. Assuming that other factors such as expectations of capital gains are not involved, which one of the following is the best approximation of the bond's current market price? Explain your answer.

A £105.62
B £106.00
C £110.00
D £100.00
E £166.00

A closer look at capital markets

Capital markets are made up of two parts: the *new-issues market* (or *primary market*) and the *second-hand market* (or *secondary market*). In the UK, the London Stock Exchange (LSE) functions as the main secondary market. There are, however, other secondary markets. These include the Alternative Investment Market (AIM), which is run by the LSE, mainly for small PLCs, and second-hand trading, which takes place online and is outside the LSE's control. Many students confuse the capital market with the stock exchange. The LSE is indeed an important part of the capital market, but it is only a part, and the terms 'capital market' and 'stock exchange' are not interchangeable. On stock exchanges, previously issued shares and bonds are sold second-hand.

The relationship between the primary and the secondary parts of the main capital market is illustrated in Figure 12.5. The actual raising of new capital or long-term finance takes place in the primary market when public companies (in the private sector of the economy) or the government (in the public sector) decide to issue and sell new marketable securities. Companies can borrow long term by selling corporate bonds, or they may sell an ownership stake in the company by

issuing shares or equity. When selling corporate bonds, the company extends its debt, and the purchaser of the bond becomes a creditor of the company.

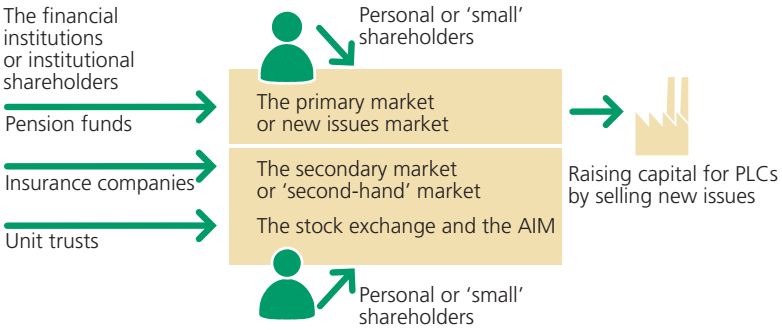


Figure 12.5 The role of capital markets in financing investment in UK industry

By contrast, new issues of shares are sold when a company 'goes public' for the first time, or when an existing public company decides to raise extra capital with a new equity issue. In the latter case, the new share issue is most often a rights issue, in which the company's existing shareholders are given the right to buy the new issue of shares at a discount.

New issues of shares are seldom sold directly on the stock exchange. Instead, the direct sale of new issues to the general public takes place in the primary market, usually being arranged by investment banks (whose functions are explained in section 12.2), via newspaper advertisements and the post. The secondary part of the capital market also performs an important economic role. Without the existence of a second-hand market, public companies or PLCs would find it difficult, if not impossible, to sell new share issues. The shares issued by private companies, which do not have a stock market listing, are generally illiquid and difficult to sell. But shares issued by listed public companies can be sold second hand on the stock exchange. The stock exchange enables shares to be converted quickly into cash. Without the stock exchange, the general public would be reluctant to buy shares that could not easily be resold. An important source of funds necessary to finance the growth of a firm would be denied to public companies.

The principal function of the London Stock Exchange and other secondary capital markets is to increase the liquidity of second-hand securities (bonds and shares), making it easier for buyers to manage their investments and sell these securities when required. This in turn makes it more likely that those with surplus funds will be willing to buy new issues of shares and bonds, thus facilitating expansion in the economic activities of PLCs and government.

STUDY TIP

Remember, equity has two very different meanings: fairness or justness, and wealth. With the latter meaning, shares are often called equities, and the market in shares is often called the equity market.

Foreign exchange markets

Over the last 60 years or so, foreign exchange markets have become more and more important in facilitating the growth of international trade and capital movements between countries. Arguably, the growth of international trade has been the main driver of economic growth in the world as a whole.

Foreign exchange markets are financial markets in which different currencies are bought and sold. International trade means that exporters and importers need to convert the funds they use to finance trade from one currency to another: for example, the UK pound sterling into euros, dollars, yen, rupees etc. Foreign exchange can be traded on either the *spot market* or the *forward market*. Spot transactions involve the immediate exchange of foreign currency, whereas forward markets involve the exchange of foreign currencies at some specified time in the future. Forward markets are used by, for example, exporters and importers to protect themselves against exchange rate risks.

SYNOPTIC LINK

Chapter 15, section 15.1, explains how foreign exchange markets work. We will see how, in the modern globalised economy, the vast majority of foreign exchange deals finance international capital investment rather than payments for exports and imports, though international trade has itself grown significantly.

TEST YOURSELF 12.12

Distinguish between a capital market and a foreign exchange market.

SECTION 12.1 SUMMARY

- Money is best defined by its two principal functions, as a medium of exchange and a store of value or wealth.
- Virtually all modern money is token money with no intrinsic value.
- Cash (notes and coins) is largely issued by the Bank of England, an agency of the state.
- Cash is the 'small change' of the system and bank deposits are by far the largest part of modern money.
- Money markets, capital markets and foreign exchange markets are examples of financial markets.
- There is an inverse relationship between bond prices and interest rates.

12.2 Commercial banks and investment banks

The difference between a commercial bank and an investment bank

The main aim of almost all banks is to make profits for their owners. The Bank of England, the country's central bank (see section 12.3), though highly profitable, is an exception, since its aims are primarily to oversee the financial system and to implement the country's monetary policy. The most commonly used way of classifying banks other than the central bank is into commercial banks and investment banks.

Commercial banks

A **commercial bank** is often also known as a **retail bank** or a **'high street' bank**. Examples include Barclays and HSBC (in the UK), Deutsche Bank (in

KEY TERM

commercial bank (also known as a **retail bank** or **high street bank**) a financial institution which aims to make profits by selling banking services to its customers.

Germany), Crédit Agricole (in France), Credit Suisse (in Switzerland) and Citibank (in the USA). Commercial banks' customers are ordinary members of the general public and businesses. To make themselves attractive to customers, commercial banks built up networks of branch banks, located in high streets and other shopping centres. However, with the development of electronic internet banking and increasing customer access to accounts on computers and smartphones, most of the large commercial banks have been closing many of their branch banks.

In summary, commercial banks are commercially run financial institutions that:

- accept deposits from the general public that can be transferred by cheque, by debit card or through an online transfer of funds undertaken
- create deposits which are lent to customers who wish to borrow from their bank.

TEST YOURSELF 12.13
Distinguish between a retail bank such as HSBC and a central bank such as the USA's Federal Reserve Bank.

Investment banks

KEY TERM
investment bank a bank which does not generally accept deposits from ordinary members of the general public. Traditional 'investment banking' refers to financial advisory work, such as advising private companies on how to become public companies by floating on the stock market (making shares available for the public to buy or sell), or advising public companies on how to buy up other companies. Investment banks also deal directly in financial markets for their own account.

Until fairly recently, **investment banks** were commonly known as merchant banks. Before the development of the internet, investment banks were clustered in the square mile of the City of London. This was partly because investment banks do not generally deal directly with ordinary members of the general public, so unlike commercial banks they do not need expensive branch networks.

Investment banks help companies, other financial institutions and other organisations (such as the government and its agencies) to raise finance by selling shares or bonds to investors and to secure against risks. They do this by underwriting share issues: that is, in return for substantial fees, investment banks guarantee to buy up all unsold shares in the event of the unsuccessful launch of a new share issue. Over recent decades, investment banks have earned large amounts of money, paid ultimately by taxpayers, through underwriting the privatisation of previously state-owned businesses, most recently the Royal Mail in 2013 (with the final 30% of share capital held by the government sold in 2015).

In addition, investment banks trade on their own behalf in shares, bonds and other financial assets. This trade takes place with other investment banks and with other financial institutions such as insurance companies and pension funds.

Investment banks are essentially global banks, the largest of which trade in financial assets which are greater in size than the GDPs of most of the world's countries. Many of the largest investment banks have been formed through the merger of a retail bank and a previously smaller independent investment bank. A good example is provided by S.G. Warburg & Co., a London-based investment bank. Warburg was listed on the London Stock Exchange and was once a constituent of the FTSE 100 index (i.e. it was one of the biggest 100 public limited companies listed on the London Stock Exchange). The bank was

Table 12.2 Top ten global investment banks measured by fees at 2019

Investment banks
J.P. Morgan
Goldman Sachs
Morgan Stanley
Bofa securities
Citigroup
Credit Suisse
Barclays
Deutsche Bank
Wells Fargo
RBC Capital Markets

Source: advratings.com

acquired by the Swiss Bank Corporation in 1995 and ultimately became part of another Swiss investment bank, UBS.

Only two of the 'big three' US investment banks, J.P. Morgan and Goldman Sachs, have at all times remained independent of the retail banking sector. The 'top ten' global investment banks in 2019 are listed in Table 12.2.

Although investment banks have headquarters in a range of countries which include the USA, the UK, France, Germany, Switzerland and Japan, they also have offices in all the world's financial centres, including the City of London and 'City-East' — that is, Canary Wharf to the east of the City proper. Indeed, all of the large global investment banks have a presence in London. These banks therefore contribute to UK economic activity, and help support the efficient functioning of the financial system. But investment banks also bring risks to the UK financial system.

Investment banks and the problems of systemic risk

As we have noted, many banks carry out both retail and investment banking activities. However, this led to some of the problems which appeared in the financial crisis that hit the global financial system in 2007. Perhaps the major problem results from **systemic risk**. Systemic risk refers to the risk of a breakdown of the entire banking system rather than simply the failure of individual banks. This is a cascading failure caused by inter-linkages in the financial system, which may result in a severe downturn in the whole economy.

KEY TERM

systemic risk in a financial context, the risk of a breakdown of the entire financial system, caused by inter-linkages in the financial system, rather than simply the failure of an individual bank or financial institution.

It is important to distinguish between systemic and one-off risks. In contrast to a one-off shock which affects only a single bank without rippling into the rest of the banking system, systemic risk affects the entire banking system and other financial institutions as well. The consequences of a systemic financial crisis can be devastating because of the role that banks and finance play in the wider economy.

Largely in response to these difficulties, national banking authorities are in the process of introducing regulations designed to separate 'high-street' and investment banking activities. Following the publication of the final version of the Vickers Report into the activities of the UK banking sector in 2011 (which was commissioned to investigate whether another financial crisis could occur), the retail banking activities of banks operating in the UK must be *ring fenced* (separated) from their investment banking activities. This became law on 1 January 2019.

Without ring fencing, an investment bank's mergers and acquisitions (M&A) department may acquire 'inside knowledge' about a company it is advising. This 'inside knowledge' could be very useful to the bank's own trading in the company's shares. An effective ring fence, which is also known as a firewall, is an internal device implemented by the bank to separate different parts of a bank's activities.

TEST YOURSELF 12.14

How did the bank crashes which occurred during the 2008–09 recession partly result from the problem of systemic risk in the financial system?

How banks create credit and bank deposits

In contrast to cash, which is tangible and can be seen and touched, bank deposits are intangible. Customers only 'see' a bank deposit when reading the statement of a bank account, or when viewing the electronic display in a cash-dispensing machine. Bank deposits are the main form of money, and most transactions are paid for by transferring bank deposits from one account to another.

STUDY TIP

You should understand that bank deposits are the main form of money and that cash is a relatively small part of the total stock of money. If you read the second edition of the book *Where Does Money Come From?* published by the New Economics Foundation in 2014, you will see that, at the time, 97.4% of the total of the broad definition of money in the UK took the form of bank and building society deposits, together with other liquid assets, which means that only 2.6% of the broad money total was actually cash.

Many students are completely mystified about how bank deposits are created. A good starting point for developing an understanding is provided by the following scenario.

Suppose we write our signatures on a scrap of paper, together with the words: 'We promise to pay the bearer £100' ('the bearer' is the intended person or firm holding the paper). We then give you the scrap of paper (a *promissory note* or 'I owe you'), and ask you to go to a shop and buy £100 worth of goods. When the time comes to pay, you must give the scrap of paper to the shop assistant. Now we all know that, in real life, shops almost always refuse to accept such an 'I owe you'. But just suppose the shop did accept our scrap of paper, believing it could then use the note to buy goods from its suppliers. Our 'I owe you' note would have become money!

Now, while ordinary individuals cannot generally create money in the manner in which we have described, banks can. If the promissory note were to be headed 'Lloyds Bank promises to pay the bearer £100', as long as people believe that Lloyds will honour its promise, the 'I owe you' can function as money.

KEY TERM

credit when a bank makes a loan, it creates credit. The loan results in the creation of an advance, which is an asset on the bank's balance sheet, and a deposit, which is a liability of the bank.

We explain the process through which banks create **credit** and new bank deposits in the next section on 'How banks create credit'. We explain credit creation, first (rather unrealistically) in the context of a 'monopoly bank', and then, more realistically, in the context of the real-world 'multi-bank' banking system.

How banks create credit

The UK banking system, and indeed that of most other countries, is a 'multi-bank' system in which there are a large number of commercial banks such as Barclays and Lloyds competing for business in pursuit of profit. Before we explain how banks in such a system create credit, we will describe how credit would be created if there was just one commercial bank in the system. This is often called a 'monopoly' banking system.

A 'monopoly' banking system

To illustrate the credit and bank deposit creating process, we are first going to assume the following:

- There is only one bank in the economy (it has the 'monopoly').
- The bank possesses only one asset, namely cash, which it uses to meet any cash withdrawals (and running down of deposits) by customers.
- The bank decides that, to maintain confidence and always to be able to meet its customers' desire to withdraw cash from their accounts, it must always possess cash equal to 10% of total customer deposits.

Although the first assumption is unrealistic, the retail banking system taken as a whole behaves as if there is just one bank. Shortly we explain how the end result of credit and deposit creation in the 'multi-bank' system in which we live is much the same as in the 'monopoly bank' system.

The assumption that there is only one retail bank means that, unlike in a system in which there are many banks, customers cannot withdraw and transfer cash from one bank to another. The second and third assumptions taken together mean that, to be sure that the bank does not run out of money, the prudent bank chooses to operate a cash ratio in which 10% of customers' deposits in the bank are backed by cash. Remember that most transactions are settled without any withdrawal of cash and, on most days, while some customers are withdrawing cash, other customers are depositing cash with the bank. This means that the bank does not need to back all its deposits with cash.

Suppose now that a member of the general public deposits £1,000 cash in the bank. From the bank's point of view, both its assets and its liabilities increase by £1,000. The cash is the bank's asset, but the £1,000 deposit credited in the customer's name is the bank's liability, since the bank is liable to honour any cash withdrawals made by the customer. The £1,000 is recorded in the bank's balance sheet in the following way:

Assets	Liabilities
Cash: £1,000	Deposit: £1,000
Total assets: £1,000	**Total liabilities: £1,000**

TEST YOURSELF 12.15

What is the difference between an asset and a liability?

As things stand, all the bank's deposit liabilities are backed with cash (i.e. the bank's *cash ratio* is 100%). If this remained the position, the 'bank' would not be a bank at all, but a safe-deposit institution, guarding (for a fee) the cash deposited by customers.

Unlike the customer depositing cash, other customers may need to borrow from the bank. Provided they are creditworthy (i.e. they can be relied on to pay back money), the 10% cash ratio the bank has chosen to work with means that the bank is in a position to lend exactly £9,000 to these customers. This may take the form of an interest-earning advance on the assets side of the bank's

balance sheet, which is matched by a £9,000 created deposit on the liabilities side of the balance sheet:

Assets	Liabilities
Cash: £1,000 Advances: £9,000	Deposit: £1,000 Created deposits: £9,000
Total assets: £10,000	**Total liabilities: £10,000**

Both the customer who made the initial deposit of £1,000 and the customers in receipt of the advances can make payments equal to £10,000 in total on their deposits. The initial £1,000 cash deposit has enabled total deposits to be increased to £10,000.

STUDY TIP

In the real-world banking system, the cash ratio is much lower than 10% — it is closer to 5%. For illustrative purposes, we will stick with the assumption of a 10% cash ratio as it makes the calculations simpler. However, see Quantitative Skills 12.3 for a calculation involving a 5% cash ratio.

Although our assumption of a monopoly bank is completely unrealistic, it has allowed us to illustrate in a simplified way that the banking system *as a whole* can expand credit, bank deposits and hence the money supply to a multiple of the assets (in this case, cash) that back the created deposits. The 'single-stage' creation of credit and deposits is possible in a 'monopoly bank' model because there is no danger that customers who have been granted loans will make payments on their deposits payable to customers of other banks. The main constraint on deposit creation is the need for the bank to hold sufficient cash to meet likely customer withdrawals. (A further constraint is the existence of customers who actually want to borrow from the bank. If no one is willing to borrow, the bank cannot expand its deposits by giving loans.)

A 'multi-bank' banking system

However, the real-world banking system contains more than one bank. Some of the biggest are HSBC, Barclays, Lloyds and NatWest. We call this a 'multi-bank' banking system. When we drop the simplification of a monopoly bank and assume a multi-bank system similar to that in the UK, the general conclusions of our simple 'monopoly' model still hold. If the increase in cash deposits is spread over all the banks, deposits can expand to £10,000, provided every bank in the system is prepared to lend to the full extent that its chosen cash ratio (known as a liquidity ratio) allows. But if only one bank is prepared to create deposits to the full, it will begin to face demands for cash that it cannot meet, losing some of its cash reserves to other banks. This will happen when the bank's customers make payments into the accounts of customers of the other banks. However, if all banks are prepared to expand deposits to the full, payments to customers of other banks will tend to cancel out. The banking system as a whole can expand deposits to £10,000, though some banks may gain business at the expense of others.

The above models illustrate how the supply of reserve assets, such as cash, available to the banks can limit their ability to create deposits. However, the banks' ability to create credit is also limited by their need to have sufficient

capital. Since the financial crisis, central banks have imposed larger minimum required capital ratios (covered later in this chapter) on the banks and this has restricted their ability to lend.

If banks hold more capital, they are less likely to become insolvent if there is a fall in the value of their assets — for example, as a result of customers being unable to repay the money they have borrowed. This is the main reason why capital ratios are imposed.

STUDY TIP

In section 12.3, when examining the functions of a central bank (in the UK, this is the Bank of England), we explain how the central bank acts as a 'lender of last resort' to the banking system. To maintain confidence in the banks and to prevent a 'run' on an individual bank, the central bank is always prepared to supply cash at a price to a bank threatened by a sudden deposit withdrawal. However, this does not mean that the central bank will bail out a bank which has made bad investments and is making losses.

TEST YOURSELF 12.16

What is the link between a bank creating credit and a bank creating a new deposit?

QUANTITATIVE SKILLS 12.3

Worked example: calculating an increase in bank deposits

A customer makes a cash deposit of £100,000 in his bank. All the retail banks in the economy choose to keep 5% of their total assets in cash.

Calculate the maximum level of total bank deposits resulting from the deposit of £100,000 into the banking system.

A 5% cash ratio means that cash equals 0.05 of total deposits.

$$\text{total bank deposits} = \frac{\text{initial deposit}}{0.05}$$

$$= \frac{£100{,}000}{0.05}$$

$$= £2{,}000{,}000$$

The answer is £2 million. As all banks in the banking system have chosen to operate a 5% cash ratio, assuming that the banking system retains the extra cash, total bank deposits can increase to £2 million following a deposit of £100,000 into the system. Notice that £100,000 is 5% of £2 million.

The structure of a retail bank's balance sheet

So far we have assumed that retail banks possess just two assets: cash and advances to customers. The left-hand column of Table 12.3 shows that, in fact, banks possess a range of other assets.

Table 12.3 The structure of a retail bank's balance sheet

Assets	Liabilities
Cash (notes and coins)	Share capital
Balances at the Bank of England	Reserves (retained profit)
Money at call and short notice (money lent to other banks via the interbank market)	Long-term borrowing (e.g. bonds issued by the bank)
Bills (commercial bills and Treasury bills)	Short-term borrowing from money markets
Investments (e.g. holdings of government bonds (gilts) and corporate bonds)	Customers' deposits (sight deposits and time deposits)
Advances (e.g. credit or loans and mortgages extended to the bank's customers)	
Fixed assets (e.g. bank buildings and other premises)	

In reality, banks possess *reserve assets* or *liquid assets* other than the cash we have already discussed. (e.g. the second, third and fourth items in the left-hand column of Table 12.3). Nevertheless, cash is the most important reserve asset and source of liquidity in the banking system. The purpose of reserve assets is to allow banks to maintain liquidity and confidence and to meet any likely demands by customers for cash. In the event of a withdrawal of cash by customers, a bank must be able to turn reserve assets into cash without suffering a capital loss.

To be as profitable as possible, a greedy or imprudent bank might be tempted to reduce cash and other liquid assets to a minimum, but this would run the risk of illiquidity and a loss of confidence among customers in the bank's ability to meet its liabilities. As was the case when the Northern Rock bank collapsed in 2007, a 'run' on the bank could then occur, leading to a crash. A 'run' on the bank is where people believe that a bank is going to run out of cash and a panic ensues, causing customers to attempt to withdraw all they can before it does run out of cash.

TEST YOURSELF 12.17
In a bank's balance sheet, what is a reserve asset?

The liquidity–profitability trade-off

As mentioned earlier, if a bank kept all its assets in the form of cash, it would not really be a bank at all, but a safe-deposit institution. The bank's profits, if it earned any, would come solely from the fees it charged customers for guarding their valuables. The bank's cash would be completely liquid, but not very profitable.

To make a profit, a bank has to make its cash go to work. Essentially, cash acts as 'high-powered' money which allows the banks to make advances to customers, creating deposits and increasing the total stock of money. The rates of interest that retail banks charge their customers on the advances granted to them are a major source of the banks' profits. Banks might be tempted to create far too many potentially profitable advances — this would

be unwise if they possess insufficient cash to meet customers' possible cash withdrawals. These banks would be operating on too low a ratio of cash and other liquid assets to the advances they have created. If a run on the banks occurred, the banks would crash.

Potential conflicts between liquidity, profitability and security

KEY TERMS

profitability the state or condition of yielding a financial profit or gain.

security secured loans, such as mortgage loans secured against the value of property, are less risky for banks than unsecured loans.

Prudent (cautious) retail banking requires banks to operate on ratios of cash and other liquid assets to advances that maintain customers' confidence in the bank, while generating acceptable profits for the bank. In other words, prudent banking involves trading off liquidity against **profitability**. Taking the UK retail banking system as a whole, over the years, cash ratios fell below 5%. Some argue that this was a major factor in the failure of the Northern Rock and Bradford and Bingley banks during the financial crisis of 2007–08, and in the government's need to rescue much larger banks such as Lloyds TSB. Since the financial crisis, cash ratios have risen again towards 5%.

Besides trading off between profitability and liquidity, banks also have to make choices with regard to the **security** of their assets. The profitability for banks of loans granted to customers depends to a significant extent on the degree of risk attached to the loans. Non-secured loans are risky because, if a customer defaults on the loan, the bank cannot recover any money. Banks therefore charge higher interest rates, and make more profit, on unsecured than on secured loans. Mortgage loans granted to house purchasers are secured loans. If the borrower is unable to repay the loan or interest on the loan, the bank or building society which provided the mortgage can repossess the property which secures the loan. This reduces the risk involved in granting long-term mortgage loans.

Table 12.4 is similar to Table 12.3, but with the liabilities column deleted and a vertical arrow added to depict the liquidity–profitability trade-off.

Table 12.4 The liquidity–profitability trade-off

Assets	
Cash (notes and coins)	**Liquidity**
Balances at the Bank of England	↑
Money at call and short notice (money lent to other banks via the interbank market)	↓
Bills (commercial bills and Treasury bills)	
Investments (e.g. holdings of government bonds (gilts) and corporate bonds)	
Advances (e.g. credit or loans and mortgages extended to the bank's customers)	**Profitability**
Non-current assets (e.g. bank buildings and other premises)	

We have already explained why cash which is held in the tills and the cash machines that banks own is the most liquid asset of all. We have also explained why advances to customers are highly profitable, but illiquid. The other items in Table 12.4 deserve a brief explanation. The first of these is balances at the Bank of England. Just as members of the general public make payments to each other by shifting ownership of bank deposits from the person making the payment (the payer) to the person receiving the payment (the payee), so the retail banks settle debts between themselves in the same



Banks may borrow from each other to service their customers' demand for money

way. To do this, all the banks keep balances at the Bank of England. As a bank can instantly withdraw cash by running down its balance at the Bank of England, the balance is as liquid as cash.

If a retail bank is temporarily short of cash, which it needs to meet customers' demand for cash over the counter or via a cash machine, it can borrow cash from other banks. The traditional name for this is 'money borrowed at call and short notice'. Money borrowed in this way which must be paid back tomorrow (overnight money) is an example of money at call. Money borrowed for a rather longer period, say a month, is money at short notice. These days, the market on which money is lent and borrowed in this way is called the *inter-bank money market*. The money *lent* in this way is the third asset in Table 12.4, but since banks borrow from, as well as lend to, other banks, money at call and short notice figures on both sides of a bank's balance sheet — as a *liability* (money borrowed) as well as an *asset* (money lent). Since the money lent can quickly be turned into cash, money at call and short notice is a highly liquid asset for the bank making the loan to other banks.

Retail banks also like to hold a portfolio of commercial bills and Treasury bills as a liquid asset (the fourth item in Table 12.4). When first issued and sold to the retail banks (by investment banks on behalf of businesses in the case of commercial bills, and by the Debt Management Office (DMO), an executive agency of the Treasury, on behalf of the government in the case of Treasury bills), bills typically have a life of three months before they mature. A bank purchasing a newly issued bill makes a profit because it buys the bill at a discount — the purchase price is lower than the bill's maturity value.

The bank earns the bill's *discount rate*, which is, in effect, the interest rate on the bill, and the money market on which commercial and Treasury bills are bought and sold is called the *discount market*. At any point in time, retail banks possess a portfolio of bills, some of which mature today or tomorrow, some in a week's time, and some in three months' time. From the bank's point of view, its bill portfolio provides it with liquid assets which are nevertheless profitable.

This is even more evident for investments — the fifth set of assets in Table 12.4. As shares never mature and might have to be sold second hand on a falling capital market, retail banks do not generally invest in company shares or equities. However, they do invest in corporate bonds issued by private-sector PLCs, and even more so in government bonds (gilt-edged securities or gilts). When first issued, most gilts have a life of anything up to 50 years before they mature. However, because gilts are issued by the government and their maturity value and yearly interest payment (the gilt's coupon) are guaranteed, banks usually regard gilts as absolutely safe and profitable investments. For the banks which own them, gilts are profitable, but illiquid in the sense that the banks may suffer capital losses if they decide to sell the gilts on a falling second-hand market. From the banks' point of view, the risk of capital loss is greater for gilts than for bills.

TEST YOURSELF 12.18
Describe the liquidity/profitability trade-off facing a commercial bank such as NatWest or TSB.

SECTION 12.2 SUMMARY

- Banks can be divided into commercial banks and investment banks, though commercial banks can be better thought of as retail banks or 'high-street' banks.
- The growing size and role of investment banks or wholesale banks, both in the UK and in other economies, contributed to the global financial crisis which started in 2007.
- Investment banks are very difficult for national governments and central banks to control; some banks that conducted both retail and investment banking activities were deemed 'too big to fail'.
- By creating credit, giving loans, banks also create deposits, which are the main constituent of the money supply.
- The amount of credit a bank can create is affected by its liquidity and its capital.
- For banks, there are potential conflicts between liquidity, profitability and security.

12.3 Central banks and monetary policy

The main functions of a central bank

KEY TERM
central bank a national bank that provides financial and banking services for its country's government and banking system, as well as implementing the government's monetary policy and issuing currency. The Bank of England is the UK's central bank. The European Central Bank (ECB) implements monetary policy for countries using the euro as their official national currency.

As we have seen, most banks are either commercial or investment banks, or both, whose main aim is to make a profit for their owners. Although it makes a considerable profit, the most significant exception is the Bank of England, which is the UK's **central bank**. For most of the period since its formation in 1694, the Bank of England was a private enterprise company. However, the Bank was nationalised in 1946 and its surplus profit goes to the state. In 1997 the Bank was given 'operational independence', which meant it was freed from political control in setting monetary policy. However, it is still 'owned' by the UK government.

A central bank has two key functions: to help its country's government maintain macroeconomic stability and to bring about financial stability in the monetary system. With regard to macroeconomic stability, the Bank of England's remit is to deliver price stability and, subject to that, to support the

government's economic objectives including those for growth and employment. The price stability remit is defined, not in terms of zero inflation, but by the government's inflation target, currently 2% on the consumer prices index (CPI). The remit emphasises the importance of price stability in achieving macroeconomic stability, and in providing the right conditions for sustainable growth in output and employment.

STUDY TIP

It is essential to understand the meaning of 'macroeconomic stability' and of a related term, 'macroeconomic performance'.

- Macroeconomic stability means stable growth, stable employment, a stable price level, and stability in the current account of the balance of payments.
- Macroeconomic performance is the extent to which these desirable objectives are being achieved.

Financial stability can be achieved, in part, through the central bank acting as lender of last resort to the banking system, and also, in part, by the central bank's monitoring and regulation of the financial system. The lender of last resort function is a standard function of central banks worldwide. It is commonly defined as the readiness of the central bank to extend loans to banks that are solvent but have short-term liquidity problems. By providing these funds, though at a price, the central bank aims to protect depositors and in the extreme case to prevent a systemic crisis in the financial system.

Central banks also carry out other related functions such as: controlling the note issue, acting as the bankers' bank, acting as the government's bank, buying and selling currencies to influence the exchange rate, and liaising with overseas central banks and international organisations.

We explained on page 440 how and why banks keep balances at the Bank of England, just as members of the general public keep balances in the retail banks. This is the 'bankers' bank' function of the Bank of England. The government also has bank accounts at the Bank of England. However, the Bank's role as banker to the government has been significantly reduced; since 1998 the Debt Management Office (DMO) has issued gilts on behalf of the Treasury, and in 2000 the DMO took over responsibility for issuing Treasury bills and managing the government's short-term cash needs.

TEST YOURSELF 12.19

With an example of each, distinguish between a central bank and an investment bank.

Implementing monetary policy

Complementing the central bank functions explained above is the Bank of England's function of implementing monetary policy on behalf of the UK government.

Monetary policy is the part of economic policy that attempts to achieve the government's macroeconomic objectives using monetary instruments, such as controls over bank lending and the rate of interest. Before 1997, monetary policy was implemented by the Bank of England on a day-to-day basis, but the Bank 'consulted' with the Treasury, which had the final say over changes in interest rates. The Treasury, which is a part of central government, and the Bank of England were then known as the monetary authorities. But the Treasury abandoned its hands-on role in implementing monetary policy in 1997 when the government made the Bank of England operationally independent and the only monetary authority.

Objectives and instruments of monetary policy

To understand monetary policy, it is useful to distinguish between policy objectives and policy instruments.

- A policy *objective* is the target that the Bank of England aims to hit.
- A policy *instrument* is the tool of control used to try to achieve the objective.

Monetary policy objectives and instruments can be classified in different ways. Policy objectives can be divided into ultimate and intermediate objectives. Policy instruments separate into those that directly affect the *supply* of new deposits that the commercial banks can create and those that influence the creation of bank deposits by affecting the *demand* for loans or credit.

TEST YOURSELF 12.20
What is the difference between a policy objective and a policy instrument?

KEY TERMS

Monetary Policy Committee a group of Bank of England officials and independent experts who meet to set interest rates in the UK eight times a year.

monetary policy instruments tools such as Bank Rate which are used to try to achieve monetary policy objectives.

Bank Rate the rate of interest that the Bank of England pays to commercial banks on their deposits held at the Bank of England (it is also referred to as the base rate).

Monetary policy *objectives* and the role of the MPC

For over 35 years, control of inflation has been the main objective of UK monetary policy. The government needs to control inflation in order to create conditions in which the ultimate policy objective of improved economic welfare can be attained. With this in mind, the government, in the person of the chancellor of the exchequer, sets the inflation rate target, which since 2003 has been 2% CPI inflation (with a margin of error of plus or minus 1%, i.e. inflation is on target if it is between 1% and 3%). The chancellor then instructs the Bank of England's **Monetary Policy Committee** (MPC) to operate monetary policy so as to 'hit' this target.

The MPC was set up after 1997 to set interest rates at the level that achieves the government's inflation target. Eight times a year, the members of the MPC (which consist of various Bank of England officials — including the governor — and various independent experts, but the composition changes over time) meet and decide whether to increase or decrease interest rates, or leave them at their current level. The MPC also sets other aspects of monetary policy, such as quantitative easing.

Monetary policy *instruments* and the role of the MPC

Monetary policy instruments are the tools used to achieve the objectives of monetary policy. In the UK, they can involve the Bank of England taking action to influence interest rates, the supply of money and credit, and the exchange rate.

Before 2009, UK monetary policy relied almost exclusively on the Monetary Policy Committee's use of the **Bank Rate**, which is the Bank of England's key interest rate, to manage the demand for bank loans. Since 2009, quantitative easing (see page 449) and other aspects of monetary policy have become increasingly important.

Managing the demand for credit through the use of Bank Rate

In normal circumstances, monetary policy, rather than fiscal policy, is now used to manage the level of aggregate demand in the economy. To understand how monetary policy is used in this way, it is worth restating the aggregate demand equation:

$$AD = C + I + G + (X - M)$$

Whereas fiscal policy can affect aggregate demand by changing the level of government spending (G), monetary policy affects the other components of aggregate demand, C, I and $(X - M)$.

Factors considered by the Monetary Policy Committee when setting Bank Rate

When deciding whether to raise or lower Bank Rate, or indeed to leave the rate unchanged, the MPC takes account of the current rate of inflation, the existence of inflationary or deflationary pressure, consumer and business confidence, the economy's growth rate, the levels of employment and unemployment, and a number of other economic indicators.

The committee is currently given a leeway of 1% above and below the 2% CPI inflation rate, but if the inflation rate moves outside this band of flexibility, the governor of the Bank of England must write an open letter to the chancellor explaining why this has happened.

TEST YOURSELF 12.21
What is Bank Rate at the time you are reading this chapter?

For several years before the 2008–09 recession, the objectives of monetary policy had been symmetrical in the sense that the Bank of England stimulated aggregate demand, which generally raises the rate of inflation, whenever inflation was expected to fall below the target rate, and tried to decrease aggregate demand to reduce inflation whenever inflation was expected to exceed the target rate.

KEY TERM
contractionary monetary policy uses higher interest rates to decrease aggregate demand and to shift the *AD* curve to the left.

Although the Bank's primary objective is price stability, it must also support the government's economic policy objectives, including those for growth and employment. Therefore, during the 2008–09 recession, negative economic growth and growing unemployment led to a situation in which controlling inflation as the main monetary policy objective was temporarily placed on the back-burner, with monetary policy being used instead to try to increase aggregate demand and bring about recovery from recession.

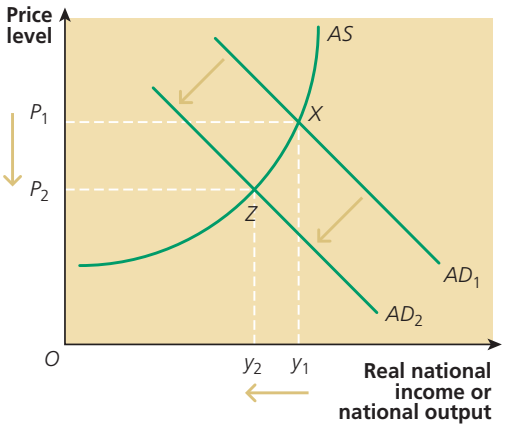


Figure 12.6 How an increase in interest rates in a contractionary monetary policy cause the *AD* curve to shift to the left

Contractionary monetary policy

In **contractionary monetary policy,** interest rates are increased so as to take demand out of the economy. In Figure 12.6, higher interest rates shift the *AD* curve to the left. However, the extent to which the price level then falls (or, more realistically, the rate of inflation falls), and/or real output falls, depends on the shape of the economy's *AS* curve. In Figure 12.6, the leftward shift of aggregate demand from AD_1 to AD_2 causes both the price level and real output to fall, respectively from P_1 to P_2, and from y_1 to y_2. This illustrates the possibility that a contractionary monetary policy, which aims to control the rate of inflation in the economy, might also cause the economy to sink into a recession. This is especially likely if the contractionary monetary policy triggers a large multiplier that shifts the *AD* curve even further to the left.

TEST YOURSELF 12.22
Explain the difference between expansionary and contractionary monetary policy.

STUDY TIP

Monetary policy shifts the *AD* curve in the economy rather than the *AS* curve. This reflects its role in the management of aggregate demand.

How an increase in interest rates decreases aggregate demand

There are three main ways in which an increase in interest rates decreases aggregate demand:

- **Higher interest rates reduce household consumption (*C*).** First, higher interest rates encourage people to save, and higher saving means that less income is therefore available for consumption. Second, the cost of household borrowing increases, which increases interest costs on mortgage and credit card debt. Borrowers have less money to spend on consumption because more of their income is being used for interest payments. Third, higher interest rates may cause asset prices, such as the prices of houses and shares, to fall. These falling prices reduce personal wealth, which reduces consumption. Fourth, falling house and share prices reduce consumer confidence, which further deflates consumption.
- **Higher interest rates reduce business investment (*I*).** Investment is the purchase of capital goods such as machines by firms. Businesses postpone or cancel investment projects because they believe that higher borrowing costs make the purchase of capital goods unprofitable. This is likely to be worsened by a fall in business confidence and increased business pessimism.
- **Higher interest rates affect exports and imports (*X* – *M*) via the exchange rate.** A higher interest rate increases the demand for pounds by attracting capital flows into the currency. The increased demand for sterling causes the pound's **exchange rate** to rise, which makes UK exports less price competitive in world markets and imports more competitive in UK markets. The UK's balance of payments on current account worsens, which shifts the *AD* curve leftwards.

KEY TERMS

exchange rate the external price of a currency, usually measured against another currency.

expansionary monetary policy uses lower interest rates to increase aggregate demand and to shift the *AD* curve to the right.

STUDY TIP

'Sterling' is the word often used for the pound in the context of the currency's role in the international economy: for example, the sterling price of imports. Economists often write about the demand for sterling.

Expansionary monetary policy

An **expansionary monetary policy**, the effect of which is illustrated in Figure 12.7, operates in the opposite way to that described above. A Bank Rate cut discourages saving, while stimulating borrowing, consumption and investment spending. Lower interest rates cause the exchange rate to fall, making exports more price competitive and imports less competitive, so exports increase. The *AD* curve shifts to the right, with the size of the shift depending on the size of the multiplier.

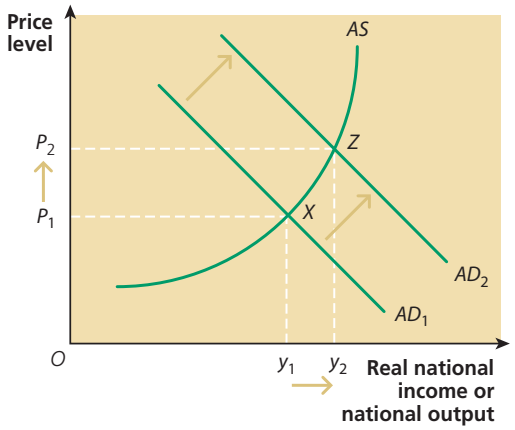


Figure 12.7 How a reduction in interest rates in an expansionary monetary policy causes the *AD* curve to shift to the right

The extent to which real output increases or the price level rises depends on the shape and slope of the economy's *AS* curve, which in turn depends on the state of the economy. When the economy produces well below the normal capacity level of output, the *AS* curve is relatively 'flat'. In these circumstances, an expansionary monetary policy is likely to increase real output (and jobs), whereas the increasing 'steepness' of the *AS* curve as normal capacity utilisation approaches means that the stimulation of real output gives way to price inflation.

The transmission mechanism of interest rate policy

The Bank of England argues that changes in Bank Rate affect aggregate demand and inflation through a number of channels, which form the transmission mechanism of Bank Rate policy. The flow chart in Figure 12.8 shows the routes through which changes in Bank Rate (the instrument of monetary policy, shown at point 1 in the diagram), eventually affect inflation (the objective of monetary policy, shown at point 11).

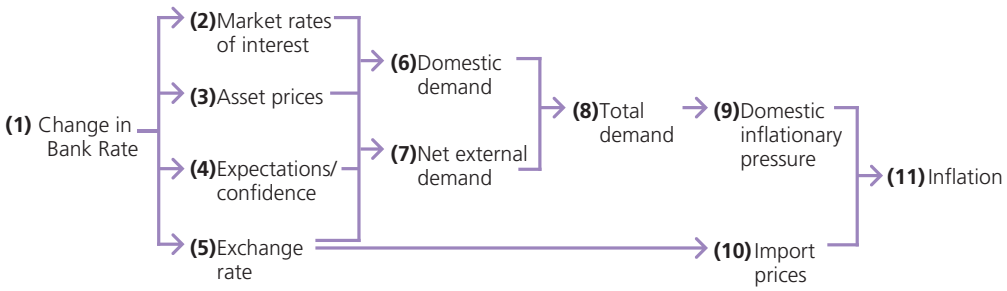


Figure 12.8 The transmission mechanism of interest rate policy

Official Bank Rate decisions (point 1 in Figure 12.8) affect market interest rates (point 2), such as mortgage rates and bank deposit rates set by commercial banks and other financial institutions. At the same time, policy actions and announcements affect expectations about the future course of the economy and the confidence with which these expectations are held (point 4). They also affect asset prices (point 3) and the exchange rate (point 5). (A rise or fall in the exchange rate also changes the prices of imports (point 10), which in turn affect the rate of inflation.)

These changes in turn affect aggregate demand in the economy (point 8). This comprises domestically generated demand (point 6) and net external demand (point 7), which is determined by export and import demand.

Domestic demand results from the spending, saving and investment behaviour of individuals and firms within the economy. Lower market interest rates increase domestic demand by encouraging consumption rather than saving by households and investment spending by firms. Conversely, higher market interest rates depress domestic spending. If Bank Rate falls, asset prices rise and people feel wealthier and generally more confident about the future. As a result, consumption increases.

Changes in aggregate demand affect domestic inflationary pressures (point 9). An increase in aggregate demand that exceeds the economy's ability to increase the supply of output creates demand-pull inflationary pressures.

Changes in the exchange rate arising from changes in interest rates also affect short-run aggregate supply. The change in import prices would lead to changes in costs of production, therefore affecting cost-push inflationary pressures.

Time lags in the transmission mechanism of interest rate policy

The Bank of England estimates a time lag of up to two years between an initial change in Bank Rate (point 1) and the resulting change in the rate of inflation (point 11). Output is affected within one year, but the fullest effect on inflation occurs after a lag of two years. In terms of the size of the effect, the Bank believes a 1% change in its official interest rate affects output by about 0.2–0.35% after about a year and inflation by around 0.2–0.4% per year after two years.

TEST YOURSELF 12.23
In relation to monetary policy, what is a transmission mechanism?

EXTENSION MATERIAL

'Conventional' and 'unconventional' monetary policy

In recent years, the monetary policy that we have described so far in this chapter, which centres on lowering or raising Bank Rate in order to manage the level of aggregate demand, has become known as 'conventional' monetary policy. In the UK, this policy worked well for most of the first decade of the twenty-first century, but broke down in the financial crash and credit squeeze of 2007 and with the onset of recession in 2008.

Figure 12.9 shows what happened to UK Bank Rate between January 2000 and May 2022. The data divide into two rather different halves. In the period before September 2008, extending back into the 1990s, Bank Rate was frequently changed, being raised or lowered in order to manage aggregate demand. However, in the months following September 2008, which was more or less the time that the recession started in the UK, Bank Rate was cut, first to a low of 0.5%, which lasted for several years, and then for a short period to 0.25%. After that, Bank Rate was raised to 0.75% before being quickly dropped to its lowest level, yet, of 0.1% in 2020 at the start of the Covid-19 pandemic. The rapidly emerging inflationary pressure of 2021 and 2022 led to a series of increases in the Bank Rate, raising it to 3.5% by December 2022 — its highest level for over 13 years.

At the time, the Bank Rate cuts from 2008 onwards were explained by some as a rather desperate attempt by the Bank of England to inject demand into the economy through the use of 'conventional' monetary policy. However, Bank Rate cuts failed to achieve this objective, and the economy experienced zero growth until late 2009 when recession gave way to a very slow economic recovery. Likewise, the Bank Rate increases which followed signalled 'conventional' monetary policy being used again to dampen inflationary pressures which were thought at the time to be re-emerging.

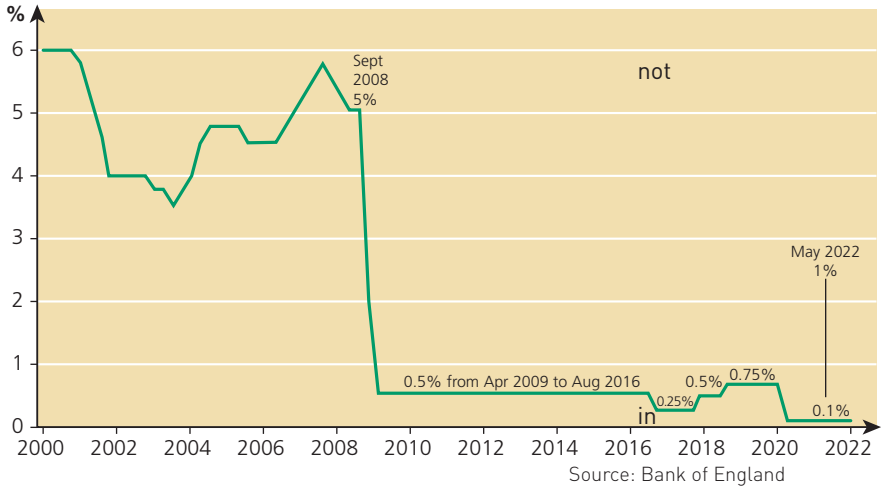
However, it was during this period that the Bank of England played down the role of 'conventional' monetary policy as a means of managing aggregate demand. Interest rate manipulation appeared not to be working, so the Bank began to use 'unconventional' monetary policy instruments to supplement the policy of keeping Bank Rate close to zero. Before we describe and explain these 'unconventional' instruments, we will first introduce two concepts which help to explain why, in a deep recession, cuts in interest rates may fail to stimulate consumer spending. These are the concepts of the zero lower bound and the liquidity trap.

The zero lower bound

When Bank Rate is cut to a very low level, at or close to 0%, there comes a point at which a floor is reached. The Bank Rate floor is called the zero lower bound (ZLB). Once the zero lower bound is reached, it becomes impossible, in the Bank of England's view, to cut Bank Rate any further, which means that the 'conventional' monetary policy of reducing interest rates to stimulate aggregate demand becomes ineffective.

The liquidity trap

Related to the zero lower bound is the liquidity trap, which describes a situation when cutting interest rates below a certain level fails to stimulate consumer spending. In a deep recession, it is likely to be the case that, however low Bank Rate is set, consumers refuse to borrow, and banks are too nervous to lend. Fear grips households and businesses, and spending in the economy dries up. A very low Bank Rate traps the country in a situation where further cuts have little or no effect on the state of aggregate demand.



Source: Bank of England

Figure 12.9 Changes in UK Bank Rate, January 2000 to May 2022

Quantitative easing

KEY TERM
quantitative easing when the Bank of England buys assets, usually government bonds, with money that the Bank has created electronically.

The main form of 'unconventional' monetary policy is called **quantitative easing** (QE). In 2009, the Bank of England's MPC began to use QE to influence commercial banks' ability to supply new loans to their customers. Quantitative easing, which is also known as the Asset Purchase Scheme, involves the Bank of England buying financial assets, usually government bonds or gilts, with new money that the Bank has created electronically. The hope is that the financial institutions (banks, pension funds and insurance companies) which sell the bonds to the Bank of England will then lend the newly created money to businesses and individuals. The latter can then invest and spend more, hopefully increasing growth.

Quantitative easing is by no means a new policy. Many monetary economists regard QE as simply a new name for an old policy called 'expansionary open market operations', which was used as far back as the 1930s.

EXTENSION MATERIAL

How quantitative easing works

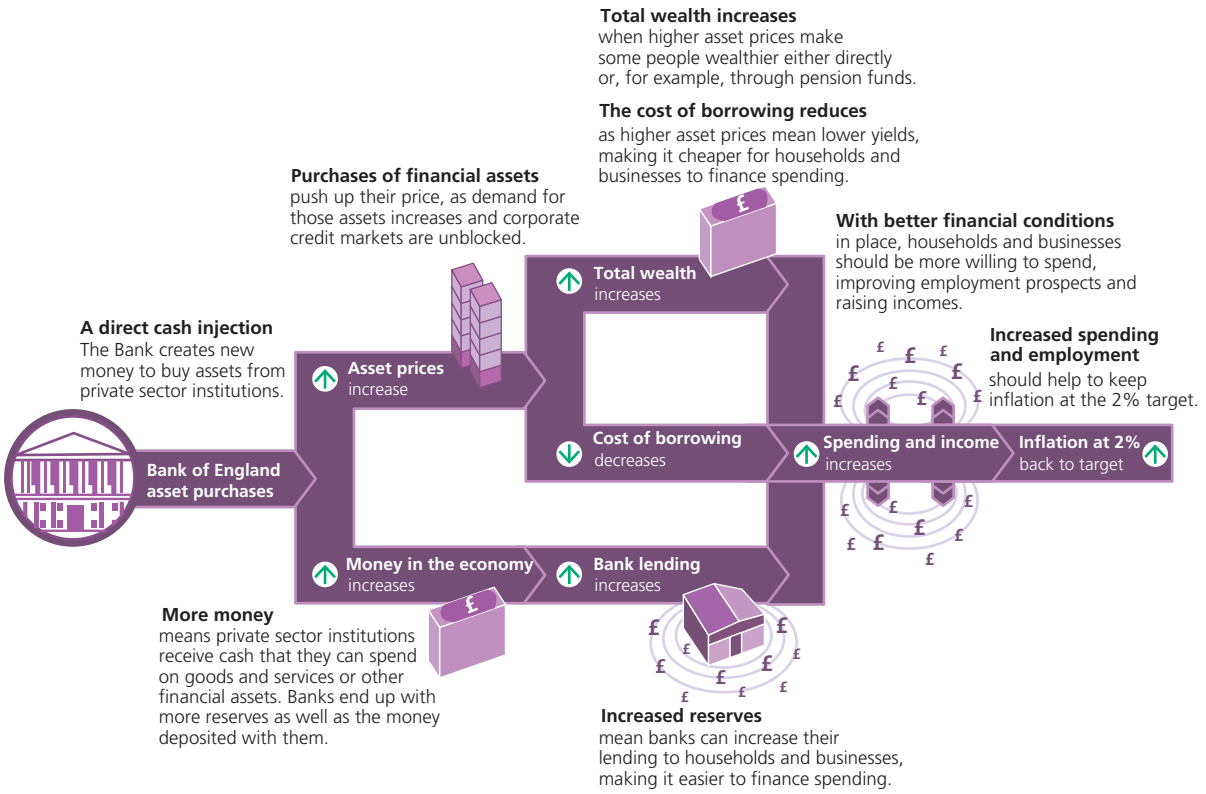
To understand quantitative easing, you should be aware that monetary policy can operate in two contrasting ways: on the demand for money, or on the supply of money. Unlike changes in Bank Rate, which attempt to influence the demand for money and, in consequence, aggregate demand, the principal aim of quantitative easing (QE) is to increase the money supply directly. When QE was started in the UK in 2009, the Bank of England hoped that the new money would be quickly spent by the households and firms who received it, thereby reviving consumer and investment spending and stimulating economic growth.

People often think that QE increases the money supply through the printing of new banknotes. Indeed, when QE started in 2009, the metaphor was used of a central bank filling a helicopter with newly printed banknotes, before dropping the 'helicopter money' on the general public, who would then spend it.

However, QE is not as simple as this. To understand why, you should appreciate that as we explained earlier in the chapter, bank deposits, and not cash, make up the majority of the money supply. QE increases the deposits that commercial banks hold at the Bank of England — and these deposits are part of the money supply. At the next stage, holding more deposits at the Bank of England increases the commercial banks' ability to lend to the general public.

The various stages through which QE proceeds are illustrated in Figure 12.10 (on the next page), published by the Bank of England. Explaining the flow chart in the diagram, the Bank wrote:

> Direct injections of money into the economy, primarily by buying gilts, can have a number of effects. The sellers of the assets have more money so may go out and spend it. That will help to boost growth. Or they may buy other assets instead, such as shares or company bonds. That will push up the prices of those assets, making the people who own them, either directly or through their pension funds, better off. So they may go out and spend more. And higher asset prices mean lower yields, which brings down the cost of borrowing for businesses and households. That should provide a further boost to spending.
>
> In addition, banks will find themselves holding more reserves. That might lead them to boost their lending to consumers and businesses. So, once again, borrowing increases and so does spending. That said, if banks are concerned about their financial health, they may prefer to hold the extra reserves without expanding lending. For this reason, the Bank of England is buying most of the assets from the wider economy rather than the banks. Normally, central banks do not intervene in private sector asset markets by buying or selling private sector debt. But in exceptional circumstances, such intervention may be warranted — for example, when corporate credit markets became blocked as the financial crisis intensified towards the end of 2008.



Source: Bank of England, 'Quantitative easing explained'

Figure 12.10 Quantitative easing's transmission route

Bank of England purchases of private sector debt can help to unblock corporate credit markets, by reassuring market participants that there is a ready buyer should they wish to sell. That should help bring down the cost of borrowing, making it easier and cheaper for companies to raise finance which they can then invest in their business.

STUDY TIP

QE is when a central bank creates money electronically which it uses to buy financial assets, mainly bonds. It is designed to support, but not to replace, Bank Rate policy.

QE1, QE2, QE3, QE4 and QE5

Quantitative easing was first used by the Bank of England in March 2009, though the USA had begun QE a few months earlier (and Japan had used QE earlier to stimulate its own economy between 2001 and 2006). In the UK there were initially three periods of quantitative easing, first in March 2009 (QE1), then in October 2011 (QE2) and finally in February and July 2012 (QE3), ending in November 2012. In the first three periods, £200 billion, then £75 billion and finally £100 billion of 'new money' was unleashed into the banking system, making a total of £375 billion.

In August 2016, the Bank of England said it would buy an additional £60 billion of UK government bonds (plus £10 billion of private-sector corporate bonds), to address uncertainty over Brexit and worries about productivity and economic growth. Some economists called this QE4.

In March 2020, an emergency meeting on the response to the emerging Covid-19 pandemic led to the Bank announcing £200 billion of purchases of

government bonds. As the pandemic continued, a further £100 billion was announced in June 2020 and then another £150 billion in November 2020. This meant that a total of £895 billion of 'new money' had been released into the UK economy (see Table 12.5).

Table 12.5 The five phases of quantitative easing in the UK between March 2009 and November 2020

QE1	£200 billion between March 2009 and January 2010
QE2	£125 billion between October 2011 and May 2012
QE3	£50 billion between July 2012 and October 2012
QE4	£60 billion from August 2016 + £10 billion purchase of corporate bonds from the private sector
QE5	£450 billion between March 2020 and November 2020
Total amount	£895 billion, including corporate bond purchases

In recent years there has been some talk of the Bank of England reversing or 'unwinding' quantitative easing by selling back to the public some of the government bonds it has accumulated while engaging in expansionary quantitative easing. This would be a form of contractionary monetary policy, taking demand out of the economy. However, with the uncertainty brought about by the Covid-19 pandemic, the war in Ukraine and the cost-of-living crisis of 2022, a reversal of QE has not yet taken place.

Has quantitative easing been successful?

It can be argued that QE1 worked in the sense that it prevented the 2008–09 recession from developing — as a result of collapsing aggregate demand — into a full-blown depression. It is debatable, however, whether the later periods of quantitative easing had much effect in bringing about recovery from recession. From 2010 onwards, the growth rate in the UK economy remained unchanged, staying very close to zero. When in 2013 a more substantial recovery did occur, further QE was not being implemented in the UK — at least until the time of the Brexit referendum in 2016.

It was also the case that, until early 2015, the continuation of QE in the USA stimulated growth in other countries, including the UK. (It is worth noting that QE was also started by the European Central Bank in 2015 to try to offset the danger of negative growth returning to the eurozone economies.)

It could be argued that QE, along with 0.5% Bank Rate, was partly responsible for the rate of inflation being higher than it otherwise would have been. It was also the case that the increased bank lending made possible by quantitative easing went, not into cheap loans to small businesses and households, but into speculative activities undertaken by the banks themselves and by the big corporations which have most benefited from the cheap money moving around those economies which implemented QE. House and land prices also went up. Asset-owning, wealthy people did much better out of quantitative easing than those who were less wealthy and who owned fewer assets.

By reducing long-term interest rates and raising bond prices, QE also benefited the UK government. This was because QE made it much cheaper for central government to borrow to finance its large budget deficit and to pay interest on the national debt. House buyers have also benefited from the cheap mortgage loans made possible by the 0.5% Bank Rate and quantitative

easing. On the other hand, savers and workers contributing to private pension schemes have suffered, although recent government policy has tried to counter this.

The *Bank of England Quarterly Bulletin* published in quarter 4, 2012 estimated that the £375 billion of QE up to that point had increased the supply of broad money in the UK by about £220 billion, or 15%, which in turn increased nominal GDP by nearly 6%. If this is true, then quantitative easing was indeed a factor causing recovery from recession. However, in an article titled 'Monetary impact of UK QE much smaller than claimed by BoE', which was published in January 2013, Simon Ward, the chief economist at Henderson Global Investors, argued that the first £375 billion of QE had delivered a monetary boost of only £78 billion.

TEST YOURSELF 12.24
Explain the difference between Bank Rate's transmission mechanism and QE's transmission mechanism.

EXTENSION MATERIAL

The Brexit vote and QE4

Immediately after the result of the Brexit referendum in 2016, the MPC acted quickly to stimulate the economy. In the following weeks, the Bank of England set out a package to prevent the economy from falling back into recession.

- First, Bank Rate was cut from 0.5% to 0.25% in the hope that ultra-low interest rates would ease borrowing costs.
- Second, the QE programme was restarted (QE4). This had two elements: £60 billion was made available to buy government bonds (a continuation of the QE 1, 2 and 3 approach); and £10 billion was allocated to buying corporate bonds to make it easier for business to raise finance for investment in the capital markets.
- Finally, the Bank introduced a new Term Funding Scheme which saw the Bank of England provide up to £100 billion of funding for commercial banks at interest rates close to 0.25%. The scheme was designed to make sure that commercial banks kept on lending to businesses, and if they reduced their lending, the rate of interest charged by the central bank would increase.

It is impossible to know what would have happened to the UK economy if the Bank of England had not taken swift action. Supporters of Brexit were highly critical of the then governor of the Bank of England, Mark Carney, and of the mainstream economics profession, arguing that the warnings of economic catastrophe that would follow a vote to leave the European Union were intended to scare the British people to vote to remain inside the EU. They claimed that predictions of immediate economic collapse were incorrect for two reasons.

- First, if the majority of those who voted, did vote to leave the EU, it would be irrational for them to start hoarding cash in fear of an economic collapse.
- Second, after a volatile couple of days, the markets settled down as businesses and traders started to realise that the UK was going to remain a member of the single market for at least another two years and that the nature of the UK's departure was yet to be negotiated.

Nevertheless, what is clear is that in the two years following the Brexit vote in 2016 the UK's GDP growth slowed significantly. Due to the uncertainty surrounding the nature of future trade relationships, new business investment slowed. The 15% depreciation in sterling against the US dollar led to an increase in inflation, and higher prices of imported goods, mostly food and energy, hit households' standards of living. Consumer spending remained strong but much of this was financed by households taking advantage of the ultra-low interest rates and reducing their savings. Economists estimate that the UK's economy was 2% smaller than it would have been had the British people voted to remain. The supporters of Brexit dismiss these economists as part of a discredited profession that wrongly predicted the breakdown of the UK economy on 24 June 2016.

APPLICATION OF ECONOMICS IN THE REAL WORLD 12.3

The inquiry into QE

In 2021 the Economics Affairs Committee of the UK Parliament held an inquiry into the effects of the quantitative easing (QE) programme. This inquiry was held because it was felt that, given the QE programme had been in place for over ten years, it had not been subject to sufficient scrutiny in Parliament.

The inquiry focused on the extent to which it had achieved its stated objectives and the effect it had had on the real economy, growth and inflation. A consideration of the risks of the programme was also included within the inquiry.

The chair of the committee stated that 'The Bank of England has become addicted to QE ... and it appears to be the answer to all the country's economic problems.' By the end of 2021, the Bank will hold bonds equivalent to 40% of UK GDP. The key findings of the committee were as follows:

- The Bank of England is widely perceived as using QE to finance the government's debt (during the Covid-19 pandemic). If this were to continue then the monetary policy of the Bank would lose credibility, making it harder to control inflation and maintain financial stability.
- Not enough information about QE is published for the public and Parliament to hold the Bank to account for its decisions. The Bank should be more open about the need for, and effects of, QE.
- No plan had been presented on how to reverse QE (known by some as 'quantitative tightening'). This should be set out, especially as inflation in 2021 was expected to rise thereafter.

The rise in inflation in 2022 and the expected slowdown in economic growth (and possible recession) reduced the urgency for quantitative tightening ... for now.

Follow-up questions

1 Why does the expected slowdown in economic growth reduce the urgency for quantitative tightening?
2 Why does continued use of QE risk the Bank losing its credibility?
3 If the Bank holds government bonds equivalent to 40% of UK GDP, couldn't the Bank just cancel these and save the government money in interest and repayments?

KEY TERM

forward guidance attempts to send signals to financial markets, businesses and individuals, about the Bank of England's interest rate policy in the months and years ahead, so that economic agents are not surprised by a sudden and unexpected change in policy.

Forward guidance

A second 'unconventional' part of monetary policy became known as **forward guidance**. The policy, which was introduced by Mark Carney, the former governor of the Bank of England shortly after taking up his appointment as governor, attempted to send signals to financial markets, businesses and individuals about the Bank of England's interest rate policy in the months and years ahead.

An aim of forward guidance was to increase the credibility of monetary policy. By using forward guidance, the Bank of England aims to calm uncertainty in otherwise uneasy financial markets. If, for example, markets fear that low interest rates will move higher, interest rates on bonds and other credit instruments and agreements will rise and begin to discourage people and companies from taking out loans or spending money. But if the Bank of England signals that it intends to keep rates low, the Bank hopes to engineer an outcome in which interest rates are indeed kept low.

And by altering expectations favourably, forward guidance aims to improve the credibility of monetary policy, thereby enabling households and business to feel calmer about their future economic prospects. According to the Bank, forward guidance means companies and mortgage borrowers can estimate for how long low interest rates will be around. If this is achieved, forward guidance is a way of converting low short-term interest rates into lower long-term interest rates.

EXTENSION MATERIAL

Forward guidance and the need for accurate economic forecasting

Mark Carney received considerable criticism for his attempts to pre-warn households, business and markets that changes in monetary policy were likely. It was felt that these hints sent confusing signals to the markets and increased the level of uncertainty.

When first announcing forward guidance in August 2013, Mark Carney said that the Bank would not consider raising Bank Rate from its low of 0.5% until the unemployment rate fell to 7% or below, which at the time the Bank forecast would happen in 2016. The forecast was wrong and unemployment fell much faster than predicted. But the MPC did not increase Bank Rate despite the rate of unemployment falling below 7% in early 2014.

If financial markets believed what Carney and the MPC were saying, the hope was that the markets would remain calm and orderly, and behave in a way which was consistent with the forward guidance strategy. There were, however, two dangers:

- First, the strategy could be knocked off course by an event such as a house price speculative bubble hitting the UK economy (or, as happened, a faster than expected fall in unemployment).
- Second, and related to this, financial markets might form their own expectations of what they thought was going to happen in the future. If markets believed that a bubble was going to occur, then, whatever the Bank of England's official policy, market operators might ignore the forward guidance strategy and raise interest rates anyway.

Either way, if traders in financial markets perceived the strategy to be 'wishful thinking', forward guidance could damage rather than increase the credibility of the Bank of England in its management of the UK economy.

Forward guidance also loses credibility if the Bank of England repeatedly 'moves the goal posts' — that is, changes policy every few months in the light of unexpected events hitting the economy. This happened for the first time in January 2014, when UK borrowing costs shot up in expectation of a rise in Bank Rate, after the fall in unemployment had proved larger than Carney had expected. A few days later, Carney signalled the end of his forward guidance linking interest rates to the unemployment rate, adding that the UK economy was in a different place from where it had been in the previous summer. Thus, the Bank of England had overhauled its forward guidance strategy less than six months after it was first implemented.

Carney said that, instead of just considering the unemployment rate, the next phase of forward guidance would be determined by a range of 18 different indicators, including productivity and the size of the output gap — the gap between potential and actual output. The Bank argued that a lack of inflationary pressure, spare capacity and 'headwinds' (a slowing down of economic progress) at home and abroad meant that 'Bank Rate may need to remain at low levels for some time to come'.

Seeking to reassure businesses and households, the Bank's Monetary Policy Committee said that when rates did eventually go up, they would do so only gradually, settling around 2–3% — below the pre-crisis norm of around 5%.

Economists employed in the financial services industry were less than convinced. For example, Schroders' UK economist branded the new guidance as a 'bamboozling cluster bomb' which, on the one hand, highlighted the degree of complexity in setting monetary policy, while, on the other hand, thoroughly confusing the average person. For cynics, forward guidance has introduced a new uncertainty into expectations on the future path of interest rates.

In general, the MPC is caught between trying to formulate monetary policy in a transparent manner that offers reassurance to markets, and the need to maintain short-term policy flexibility. Each month the MPC examines a variety of economic data and forecasts which can easily change given the unpredictable macroeconomic climate of recent years. It is extremely difficult for economic models to predict how monetary policy decisions may affect the economy. Rather than attempting to give forward guidance and creating confusion, it might be better for monetary policy-makers to be clearer about how they will react to a given set of circumstances as they unfold.

The current governor of the Bank of England — Andrew Bailey — was appointed to the position in 2020. He faced criticism over comments he made that some people thought were forward guidance but which were later clarified as merely 'commentary' on likely events. Bailey had led some to believe that interest rates would rise in November 2021 when he made comments about the risk of inflationary expectations and how the Bank might need to take action (i.e. raise interest rates). When interest rates did not rise as some expected, he made a statement about how the difference between 'guidance' and 'commentary' was unclear. Many now understand this clarification to mean that forward guidance will no longer be used as part of the Bank's approach to monetary policy.

The relationship between changes in interest rates and the exchange rate

A fall in UK interest rates causes financial capital to flow out of the pound and into other currencies in search of better rates of return. This reduces the demand for pounds and increases the supply of pounds on the foreign exchange market, which in turn causes the pound's exchange rate to fall. Changes in export and import prices brought about by a fall in the exchange rate affect inflation in two ways.

- First, a falling exchange rate increases the prices of imported food and consumer goods. This increases the rate of inflation in the UK. At the same time, increased prices of imported raw materials and energy create cost-push inflationary pressures in the UK (import-cost-push inflation).
- Secondly, a falling exchange rate reduces UK export prices, while raising the price of imports. This feeds into the inflationary process described above, by increasing the demand for UK exports and persuading some UK residents to buy more home-produced goods and fewer imports. This adds to demand-pull inflationary pressures. It is also likely to improve the UK's balance of payments on current account.

TEST YOURSELF 12.25

If the nominal interest rate is 6% and rate of inflation is 4%, what is the real rate of interest?

QUANTITATIVE SKILLS 12.4

Worked example: calculating real Bank Rate

In May 2022, Bank Rate was 1.0% and CPI inflation was 7.0%. What was Bank Rate expressed in real terms?

real rate of interest = nominal rate of interest – rate of inflation

In May 2022, the Bank Rate expressed in real terms was 1.0% – 7.0%, which is –6%.

SECTION 12.3 SUMMARY

- The main functions of the Bank of England are to implement the government's monetary policy and maintain the stability of financial markets.
- In recent decades, Bank Rate has been the main monetary policy instrument, aimed primarily at targeting the rate of inflation — that is, hitting a target of a 2% CPI rate of inflation. The target is set by the government and not the Bank.
- In an *AD/AS* diagram, monetary policy is illustrated by shifts of the *AD* curve to the right (expansionary monetary policy) and to the left (contractionary monetary policy).
- A 'good' deflation, caused by falling commodity prices and effective supply-side policies, might be deemed a success, but a 'bad' deflation, caused by a collapse of aggregate demand, would be a policy failure.
- In 2009, monetary policy switched away from relying solely on Bank Rate changes to a policy of quantitative easing (QE), through which the Bank of England buys bonds, reducing long-term interest rates and increasing the liquidity of financial markets.
- As well as QE, other 'unconventional' monetary policy instruments have been introduced, notably forward guidance, in 2013.

12.4 The regulation of the financial system

The meaning of regulation

We first explained regulation in Chapter 8, section 8.8, when we said that governments use regulation to try to correct market failures and to achieve a socially optimal level of production and consumption. We also explained how regulation is used to limit and deter monopoly exploitation of consumers.

Regulation means imposing rules and sometimes laws which limit the freedom of individuals and businesses to make decisions of their own free will. *Financial regulation* involves limiting the freedom of banks and other financial institutions, and of the people they employ, to behave as they otherwise might wish to do.

TEST YOURSELF 12.26
Explain how regulation differs from deregulation.

Financial regulation in the UK

The recent history of financial regulation in the UK

Before 2001 the Bank of England largely controlled financial regulation in the UK. Back in 1985 a self-regulatory board, the Securities and Investments Board, was created with a limited number of regulatory powers. But after a series of scandals in the 1990s, culminating in the collapse of Barings Bank, there was a desire to bring to an end the self-regulation of the financial services industry and to consolidate regulatory responsibilities which previously had been split among multiple regulators.

In 1997, the name of the Securities and Investments Board was changed to the Financial Services Authority (FSA), an external regulator with a range of regulatory powers.

The 1998 Bank of England Act, which also formally established the MPC and thereby radically changed the role of the Bank of England, transferred the responsibility for the supervision of deposit-taking institutions from the Bank to the FSA. However, the FSA did not start to exercise these powers until 2001 when the Bank of England was instructed to concentrate on banking, with its regulatory function taken away and given to the FSA.

But due to perceived regulatory failure of the banks during the financial crash of 2007-08, the Treasury decided to abolish the FSA. It had been criticised for failing to spot the lending boom before 2007 and the subsequent bust, and for not curbing risky trading by banks, the end result of which was to see the collapse of banks such as Northern Rock and the Royal Bank of Scotland. The regulatory function of the FSA had lasted from 2001 until 2013, when the authority was eventually wound up.

With effect from 1 April 2013, some of the FSA's responsibilities were given to a new committee, the Financial Policy Committee (FPC) at the Bank of

KEY TERM

Financial Policy Committee
the part of the Bank of England charged with the primary objective of identifying, monitoring and taking action to remove or reduce systemic risks with a view to protecting and enhancing the resilience of the UK financial system. The committee's secondary objective is to support the economic policy of the government.

England, with the others being split between two new agencies, the Prudential Regulation Authority (PRA) — itself a part of the Bank of England — and the Financial Conduct Authority (FCA), which is not part of the Bank.

The Financial Policy Committee

Through the **Financial Policy Committee** (FPC), the Bank of England once again has responsibility for the Bank's second key function: maintaining financial stability. The FPC is primarily responsible for *macro*prudential regulation whereas the PRA and FCA are mainly responsible for *micro*prudential regulation. Macroprudential regulation is concerned with identifying, monitoring and acting to remove risks that affect the stability of the financial system as a whole. Microprudential regulation focuses on ensuring the stability of individual banks and other financial institutions; it involves identifying, monitoring and managing risks that relate to individual firms.

APPLICATION OF ECONOMICS IN THE REAL WORLD 12.4

Macroprudential policy at the Bank of England

An important element of reforms to the UK financial regulation is the creation of a macroprudential authority at the Bank of England — the Financial Policy Committee (FPC). In the 2008-09 recession, policy-makers around the world recognised that focusing separately on price stability and on microprudential regulation of individual firms and markets was not enough. A broader approach — macroprudential policy — was needed to ensure the resilience and stability of the financial system.

Under the Bank of England Act 1998, as amended by the Financial Services Act 2012, the Bank has a statutory objective to protect and enhance the stability of the UK's financial system. The FPC is tasked with helping the Bank meet that objective and, subject to that, also supporting the government's economic policy, including its objectives for growth and employment.

In the period leading up to the 2007–08 financial crash, insufficient attention was paid to tackling risks and vulnerabilities across the financial system as a whole. The FPC fills that gap by identifying, monitoring and, crucially, taking action to remove or reduce systemic risks to the resilience of the financial system.

In carrying out its statutory responsibilities, the FPC is supported by the objectives of the microprudential regulator, the Prudential Regulation Authority (PRA), which is part of the Bank of England. The PRA is responsible for the microprudential regulation of individual deposit-takers, insurers and major investment banks.

The Financial Conduct Authority (FCA) is a separate institution responsible for ensuring that relevant markets function well, for conduct regulation and for microprudential regulation of financial services firms not supervised by the PRA, such as smaller traders and dealers in financial markets and independent financial advisers.

One way in which the FPC can mitigate threats to the resilience of the financial system is by raising awareness of systemic risks among financial market participants. The FPC is required to publish a Financial Stability Report twice a year, which must identify key threats to the stability of the UK financial system.

At times, simply warning about risks may be sufficient to catalyse action within the private sector to reduce vulnerabilities. But experience from before the crisis showed that warnings alone are not always enough.

Source: adapted from the *Bank of England Quarterly Bulletin*, Quarter 3, 2013

Follow-up questions

1 What is the difference between macroprudential and microprudential regulation?
2 Explain the differences between the Bank of England's Monetary Policy Committee (MPC) and the Bank's Financial Policy Committee (FPC).
3 'The Bank has a statutory objective to protect and enhance the stability of the financial system of the United Kingdom.' Describe one event which might reduce the financial stability of the United Kingdom.
4 What is a systemic risk? Give an example of a 'systemic risk to the resilience of the financial system'.

KEY TERMS

Prudential Regulation Authority the part of the Bank of England responsible for the microprudential regulation and supervision of banks, building societies, insurers and major investment firms.

Financial Conduct Authority aims to make sure that financial markets work well so that consumers get a fair deal, by ensuring that the financial industry is run with integrity and that consumers can trust that firms have their best interests at heart, and by providing consumers with appropriate financial products and services.

The Prudential Regulation Authority

The **Prudential Regulation Authority** (PRA) is responsible for the microprudential supervision of banks, building societies, credit unions, insurers and major investment firms. It sets standards and supervises financial institutions at the level of the individual firm. The authority regulates by setting standards which financial institutions are required to follow, and supervises by assessing the risks posed by individual financial firms and taking action to make sure they are managed properly. It aims to promote the soundness of banks and other firms providing financial services, so that the stability of the UK financial system is enhanced.

The PRA may require individual institutions to maintain specified capital and liquidity ratios so as to try to ensure that, if a financial firm fails, it does so in a way that avoids significant disruption to essential financial services.

The Financial Conduct Authority

The aim of the **Financial Conduct Authority** (FCA) is to regulate the financial services industry in order to:

- *protect consumers* by securing an appropriate degree of protection for them
- *protect financial markets* so as to enhance the integrity of the UK financial system
- *promote effective competition* in the interests of consumers.

Bank failures, liquidity assurance and moral hazard

A bank's capital is equal to the value of its assets minus the value of its liabilities, and if the value of its assets falls so that it runs out of capital, it is insolvent or bankrupt. A bank can also fail because it does not have sufficient liquidity. Insufficient liquidity makes a bank vulnerable to a run on the bank, which can cause the bank to fail, even if its assets are greater than its liabilities. However, the willingness of the central bank to act as lender of last resort and provide liquidity insurance increases confidence in the stability of the banks. It is necessary because the banks borrow short term (e.g. overnight money borrowed on the inter-bank market), but lend long term (e.g. mortgage loans which finance house purchase).

As the 2007 financial crash showed, banks are sometimes tempted to take too many risks in pursuing the huge profits that lending allows. They do this because they believe that the Bank of England, in its role as lender of last resort, and the government through its bailouts, will not allow banks to fail.

KEY TERM

moral hazard the tendency of individuals and firms, once protected against some contingency, to behave so as to make that contingency more likely.

This illustrates the problem of **moral hazard**. Moral hazard exists when a firm or individual pursues profit and takes on too much risk in the knowledge that, if things go wrong, someone else will bear a significant portion of the cost. Investing in high-risk assets can lead to high profits, but unless there is the possibility that financial institutions will be allowed to fail, there is insufficient incentive to act prudently.

When the Northern Rock bank got into financial difficulty in 2007, the Bank of England's initial reaction was to allow the bank to fail, as a way of teaching a lesson to other banks that might be taking excessive financial risks. However, the Bank soon realised that, instead of promoting less risky behaviour,

allowing a relatively small bank to fail could trigger a crisis within the whole commercial banking system, the effects of which would dwarf those of bailing out a reckless bank. The Bank of England quickly changed its policy and in effect nationalised the bank. Instead of allowing banks to go under, the Bank and the government started to bail out banks such as RBS, which failed in October 2008 and was then part-nationalised. Rescuing banks deemed 'too big to fail' had replaced the policy of allowing banks to fail to teach others about the dangers of excessive risk taking.

However, the Bank of England's recognition of the problem of moral hazard did lead to other more robust measures being introduced, such as imposing 'firewalls' between the retail and the investment banking activities of the banks. In principle, such measures are meant to allow the riskier parts of the banks to fail without impairing the provision of their retail banking services.

Liquidity and capital ratios

Over time, banks have failed or required government assistance either because they lacked liquidity, or because they had inadequate capital, or through a combination of both of these contributory factors. Liquidity and capital are *distinct* but *related* concepts. Each plays an essential role in understanding a bank's viability and solvency.

We can use a family's finances to illustrate the differences between liquidity and capital. On the liquidity side, money in a family's bank account and any cash the family has on hand that can be used quickly and easily to pay its bills are measures of the family's liquidity position. On the capital side, the family's assets include not just the money in its bank account, but also its home, savings accounts and other investments. The family debts, or money it owes, such as housing rent or a mortgage, are its liabilities. So the difference between the family's debts and its assets would provide a measure of the family's capital position.

KEY TERMS

liquidity ratio the ratio of a bank's cash and other liquid assets to its deposits.

capital ratio the amount of capital on a bank's balance sheet as a proportion of its loans.

We have already looked at a retail bank's **liquidity ratio** in the context of explaining how banks create credit and money. A bank's liquidity ratio is the ratio of cash and other liquid assets owned by the bank to its deposit liabilities. Liquidity is a measure of the ability and ease with which assets can be converted to cash. Liquid assets are those that can be converted into cash quickly, if needed to meet financial obligations. Examples of liquid assets are cash itself, balances held at the central bank, and holdings of short-dated government debt. To remain viable, a financial institution must have enough liquid assets to meet its short-term obligations, such as withdrawals of cash by depositors. Liquidity problems arise when a bank does not hold sufficient cash (or assets that can easily be converted into cash) to repay depositors and other creditors.

A bank's capital, which is the difference between the value of the bank's assets and its liabilities, represents the net worth of the bank or its value to those who own the bank. Alternatively, assuming the bank is a company, the bank's capital can be thought of as the shareholders' stake in the bank. Capital, which acts as a financial cushion that allows a bank to absorb unexpected losses, protects creditors in case the bank's assets fall in value. A bank's **capital ratio** is the amount of capital on a bank's balance sheet as a proportion of its loans. The ratio can be used as an indicator of a bank's financial health. Bank regulators are likely to require the capital ratio to be above a prescribed minimum level.

If banks do not have enough capital, they are at risk if the value of their assets falls. While *insufficient liquidity* makes a bank vulnerable to a run on the bank, *insufficient capital* exposes the bank to the risk of a fall in the value of its assets. If a bank invests in assets which fall in value because, for example, some customers default on their loans, this will result in losses and reduce the bank's capital. If the reduction in the value of a bank's assets is so large that it destroys the whole of the bank's capital, the bank is technically bankrupt and cannot continue trading.

In 2010 and 2011 the Basel III agreement was signed by the members of the Basel Committee on Banking Supervision, which include the Bank of England. The agreement, which is voluntary, is supposed to strengthen bank capital requirements by increasing bank liquidity and decreasing borrowing by banks. Portions of Basel III have already been implemented but the final implementation will begin in 2023 with full compliance expected by 2028.

TEST YOURSELF 12.27
Distinguish between a capital ratio and a liquidity ratio.

Systemic risk and the impact of problems that arise when financial markets affect the real economy

We defined and briefly explained systemic risk earlier in the chapter. We stated that financial systemic risk refers to the risk of a breakdown of the entire financial system, caused by inter-linkages within the financial system, rather than simply the failure of an individual bank or financial institution within the system.

The largest and most dramatic recent example of financial systems being threatened by systemic failure occurred in the so-called 'credit crunch', which began in 2007 and extended into 2009. The credit crunch, defined as 'a severe shortage of money or credit', began on 9 August 2007 when the French bank BNP Paribas told its depositors that they would not be able to take money out of two of the bank's funds because BNP Paribas could not value the assets in them, owing to a 'complete evaporation of liquidity' in the market.

This was the clearest sign yet that banks were refusing to do business with each other, which triggered a sharp rise in interest rates. However, the roots of the credit crunch were in place several years earlier.

Virtually all firms, large and small, require a reliable supply of credit or bank lending in order to remain in business. In 'normal' circumstances, the banking system provides this liquidity, which businesses, consumers and governments usually take for granted. Historically, the source of this liquidity was banks borrowing household savings, which the financial intermediaries then passed on through their lending for others to spend.

However, by 2007, banks throughout the world, but particularly in the USA, were raising the funds they lent to customers by borrowing from each other rather than from households. The funds were borrowed on the interbank market. In the USA, much of the borrowed funds was lent in the form of mortgages to low-income customers who were bad credit risks. These loans became known as *sub-prime* mortgages (in contrast to *prime* mortgages, which are secured loans granted to low-risk home owners). From a bank's point

of view, a mortgage granted to a customer is an asset. For the borrower, it is a liability, since the house owner must eventually pay back the loan and pay interest in the intervening months and years. By definition, a sub-prime mortgage is a risky asset since there is a danger of the loan turning into a bad debt, which the bank that lent the money cannot recover.



The failure of BNP Paribas to honour its financial obligations started the credit crunch

In 2007, the credit crunch developed because the banks that had created sub-prime mortgages repackaged the risky assets and sold them on to other banks as if they were prime mortgages. In essence, banks were buying 'toxic debt' from each other, without understanding that the repackaged assets were extremely risky. As banks realised that many of their so-called 'assets' were more or less worthless, the situation quickly deteriorated and the supply of liquidity began to freeze. Banks became unwilling to lend to each other because they distrusted each other's creditworthiness. At the next stage, the credit crunch triggered a financial meltdown when banks either collapsed (Lehman Brothers in the USA) or were partly or completely nationalised by governments (Royal Bank of Scotland in the UK).

The financial crisis that followed BNP Paribas's failure to honour its obligations eventually destabilised the real economies of many countries, including the USA and the UK. The financial crisis caused a collapse of aggregate demand, which in turn led to rising unemployment and recession. The credit crunch paralysed financial markets, threatening affected countries with the problem of systemic failure in their financial systems. Systemic failure was eventually largely averted through government and central bank intervention to rescue financial systems, but by this time much real economic damage had been done.

SECTION 12.4 SUMMARY

- A financial market is a market in which financial assets or securities are traded. The financial markets you need to know are money markets, capital markets and foreign exchange markets.
- In the light of what happened in the financial crisis and credit crunch after 2007, financial regulatory authorities have been set up in the UK to try to prevent systemic risk and moral hazard problems.
- Systemic risk in a financial context refers to the risk of a breakdown of the entire financial system, caused by inter-linkages within the financial system, rather than simply the failure of an individual bank or financial institution within the system.
- Moral hazard is the tendency of individuals and firms, once protected against some contingency, to behave so as to make that contingency more likely.
- The main authorities involved in financial regulation in the UK are the Bank of England's Financial Policy Committee (FPC), the Prudential Regulation Authority (PRA), which is also a part of the Bank of England, and the Financial Conduct Authority (FCA).

Questions

1. What is money? Explain how both the Bank of England and privately owned banks create money.
2. Explain how changes in Bank Rate may affect bank lending and the economy.
3. How has quantitative easing operated in the UK? Evaluate the case for and against the Bank of England extending quantitative easing.
4. Do you agree that monetary policy has been highly effective in the UK in recent years? Justify your answer.
5. Distinguish between capital markets and money markets. Explain how UK capital markets help both public limited companies (PLCs) and the government to undertake their functions.



13 Fiscal policy and supply-side policies

Fiscal policy is the part of economic policy in which the government attempts to achieve its policy objectives using government spending, taxation and the government's budgetary position (balanced budget, budget deficit or budget surplus). Supply-side economic policy is the set of government policies that aim to change the underlying structure of the economy and improve the economic performance of markets and industries, and also of individual firms and workers within markets. Supply-side fiscal policy is an important supply-side policy, but supply-side policies encompass more than just fiscal policy measures. Section 13.1 offers an in-depth explanation of certain key aspects of fiscal policy. Section 13.2, which makes up the second half of the chapter, provides a similar coverage of supply-side policies.

LEARNING OBJECTIVES

These are to understand:

- what is meant by fiscal policy and the difference between macro and micro fiscal policy
- how fiscal policy can be used to influence both aggregate demand and aggregate supply
- the types of and reasons for public expenditure and taxation, and how they can affect the pattern of economic activity
- direct and indirect taxation, and progressive, proportionate and regressive taxes
- the principles of taxation and the role and relative merits of different UK taxes
- budget deficits, surpluses and the national debt, and their consequences for macroeconomic performance
- the role of the Office for Budget Responsibility
- supply-side policies and supply-side improvements
- the difference between free-market and interventionist supply-side policies
- the role of supply-side policies in reducing the natural rate of unemployment and in achieving the government's macroeconomic objectives
- the distinction between the microeconomic and macroeconomic effect of supply-side policies.

13.1 Fiscal policy

The meaning of fiscal policy

KEY TERM
fiscal policy the use by the government of government spending and taxation to try to achieve the government's policy objectives.

Fiscal policy is the part of a government's overall economic policy that aims to achieve the government's economic objectives through the use of taxation, public spending and the government's budgetary position. As an economic term, fiscal policy is often associated with Keynesian macroeconomic theory and policy. Between the 1950s and the late 1970s, Keynesian governments used fiscal policy to manage the level of aggregate demand.

However, fiscal policy can also have microeconomic functions, especially when used to make markets more flexible and dynamic. At a micro level, government will use taxation and subsidies to influence consumer behaviour, and income tax and the welfare system to create incentives in the labour market.

How fiscal policy can be used to influence both aggregate demand and aggregate supply

KEY TERMS
balanced budget achieved when government spending equals government revenue ($G = T$).

budget deficit occurs when government spending exceeds government revenue ($G > T$).

budget surplus occurs when government spending is less than government revenue ($G < T$).

demand-side fiscal policy used to increase or decrease the level of aggregate demand (and to shift the *AD* curve right or left) through changes in government spending, taxation and the budget balance.

deficit financing deliberately running a budget deficit and then borrowing to finance the deficit.

Fiscal policy and aggregate demand

In order to explain demand-side fiscal policy, we need first to explain **balanced budgets**, **budget deficits** and **budget surpluses**. Using the symbols G for government spending and T for taxation and other sources of government revenue, the three possible budgetary positions a government can have are:

- $G = T$: balanced budget
- $G > T$: budget deficit
- $G < T$: budget surplus.

STUDY TIP
Make sure you don't confuse a *budget* deficit with a *balance of payments* deficit. They are completely different — one is concerned with the government's finances, the other with the position on net external trade.

A budget deficit occurs when public-sector spending exceeds revenue. It is important not to confuse *financing* a budget deficit with *eliminating* a budget deficit. A budget deficit can be eliminated by cutting public spending or by raising taxation, both of which can balance the budget or move it into surplus. Assuming a budget deficit persists, the extent to which spending exceeds revenue must be financed by public-sector borrowing.

During the Keynesian era, a period which extended from the 1950s until the late 1970s, fiscal policy was used primarily to manage the level of aggregate demand in the economy. Keynesian fiscal policy, or **demand-side fiscal policy**, centred on the use of **deficit financing** to inject demand into the economy.

TEST YOURSELF 13.1
Distinguish between 'demand-side' fiscal policy and 'supply-side' fiscal policy.

SYNOPTIC LINK
Pages 487–88 later in this chapter explain in detail the meanings of budget deficits and budget surpluses, such as what causes them, and their consequences for economic performance.

Deficit financing describes a situation in which the government runs a budget deficit, usually for several years, deliberately setting public-sector spending at a higher level than tax revenues and other sources of government revenue. For each of the years in which the government runs a budget deficit, the shortfall of tax revenue has to be financed through a positive borrowing requirement. The aim is to achieve full employment and to stabilise the economic cycle, without at the same time creating excessive inflationary pressures.

Keynesian fiscal policy was implemented with varying degrees of success in the decades before 1979. Since 1979, governments have increasingly relied on monetary policy for the management of the macroeconomy. However, the economic shocks caused by the financial crisis and the Covid-19 pandemic have been partially dealt with by active Keynesian fiscal policy — with large amounts of demand pumped into the economy from both tax reductions and increased government spending.

Some of the Keynesian views which strongly influenced the use of fiscal policy from the 1950s until 1979 were as follows:

- Left to itself, an unregulated market economy results in unnecessarily low economic growth, high unemployment and volatile business cycles.
- A lack of aggregate demand, caused by a tendency for the private sector to save too much and invest too little, can mean that the economy settles into an under-full-employment equilibrium characterised by demand-deficient unemployment.
- By deliberate deficit financing, the government can, using fiscal policy as a demand management instrument, inject demand and spending power into the economy to eliminate deficient aggregate demand and achieve full employment.
- Having achieved full employment, the government can then use fiscal policy in a discretionary way (i.e. changing tax rates and levels of public spending to meet new circumstances) to fine-tune the level of aggregate demand. For much of the Keynesian era, governments believed that fiscal policy could achieve full employment and stabilise the economic cycle, while avoiding an unacceptable rate of inflation.
- The overall stance of fiscal policy and, indeed, of economic policy in general, was orientated towards the demand-side of the economy. The more microeconomic elements of fiscal policy, such as transfers to industry, were aimed at improving economic performance on the supply side. But on the whole, supply-side fiscal policy was treated as subordinate to the macroeconomic management of aggregate demand and to the

assumption that output would respond to demand stimulation. In any case, the microeconomic elements of Keynesian fiscal policy were generally interventionist rather than non-interventionist, extending rather than reducing the state's role in the mixed economy.

- Central to Keynesian fiscal policy was the assumption that any increase in government spending would be affected by the multiplier and result in a proportionally larger increase in national income. A relatively large multiplier means that changing the levels of government spending, taxation and the budget deficit (or surplus) can be quite effective in managing aggregate demand. However, real-world government spending multipliers are generally thought to be small, and in most cases are not significantly different from unity (1).

SYNOPTIC LINK

Refer back to Chapter 10 and refresh your knowledge of the national income multiplier in the context of shifts of the economy's aggregate demand curve. Keynesian demand management was based on the utilisation of 'multiplier effects' on demand and growth in the economy.

EXTENSION MATERIAL

Deficit financing and Keynesian macroeconomic policy

For many decades before the advent of the Keynesian era after the Second World War, UK governments believed they had a moral duty to balance their budgets. This has been called *sound finance*. The traditional view was that a budget surplus placed the government in the moral position of a thief, stealing from taxpayers. If the government ran a budget deficit, it would be in the moral position of a bankrupt, perceived as not being able to manage its finances. Since both these budgetary positions were regarded as wrong or undesirable, the government's fiscal duty was to aim for a balanced budget.

However, in the 1930s, John Maynard Keynes established a new theory that legitimised deficit financing and overturned the view that a government should always aim to balance its budget. The dominance of Keynesian theory lasted until the late 1970s, when Keynesianism was replaced by a return to pro-free-market economics and a belief in balanced budgets. Keynes argued that mass unemployment in the Great Depression was caused by deficient aggregate demand. He believed that, in the economy as a whole, too little spending was taking place because households and firms in the private sector were saving too much and spending too little. In the 1950s and 1960s, Keynes's followers, the Keynesians, argued that, in this situation, if the government deliberately runs a budget deficit, the deficit can be financed by the government, first borrowing and then, in its public spending programme, spending the private sector's excess savings. This injects spending into the economy and (in Keynesian theory, at least) gets rid of demand-deficient unemployment.

In the UK, the Keynesian era lasted until the late 1970s. However, there were signs of a move away from Keynesian economics as early as 1976 when the Labour prime minister, James Callaghan, gave a famous speech where he described how it would no longer be an option for governments to 'spend [their] way out of a recession'. The end of the Keynesian era gave way to a return of free-market theory and policies that rejected deficit financing and the use of fiscal policy to manage the level of aggregate demand.

SYNOPTIC LINK

Refer back to the extension material in Chapter 9 on 'The growth of Keynesian economics' (page 311) for further discussion of Keynesian economics.

APPLICATION OF ECONOMICS IN THE REAL WORLD 13.1

Budget day

The UK government's financial year runs for the 12 months beginning on 5 or 6 April. In October or November of the previous year, the chancellor of the exchequer (an elected member of parliament (MP) who is responsible for the economy) reads their budget speech to the House of Commons outlining how the government proposes to collect tax revenue in the forthcoming financial year. Part of the budget speech, which is published in the 'Red Book' (formally called the Financial Statement and Budget Report (FSBR), but commonly named after the colour of its cover), contains the chancellor's analysis of the current performance of the UK economy. This is the part of the budget which is of most interest to economists.



Jeremy Hunt became chancellor of the exchequer in October 2022

The extract below has been taken from the Treasury website.

History of the budget

The word 'budget' derives from the term 'bougette' — a wallet in which either documents or money could be kept. The longest Budget speech is believed to have been by William Gladstone on 18 April 1853, lasting four hours and forty-five minutes. The Budget box or 'Gladstone box' was used to carry the Chancellor's speech from Number 11, Downing Street to the House of Commons for over 100 consecutive years. The wooden box was hand-crafted for Gladstone, lined in black satin and covered in scarlet leather.

Before going to Parliament to deliver the statement, the Chancellor holds up the red box outside Number 11 to waiting photographers. Chancellors are allowed to refresh themselves with alcoholic drinks during their Budget speech — no other Member of Parliament can do this! The last few Chancellors have opted for non-alcoholic drinks. Traditionally, the Leader of the Opposition — rather than the Shadow Chancellor — replies to the Budget speech. This is usually followed by four days of debate on the Budget Resolutions (the basic parts of the Budget that renew taxes), covering different policy areas such as health, education and defence.

Although the budget day has been in the autumn for the last few years it has also been held in the spring on a number of occasions in the last thirty years.

Follow-up questions

1 What is the purpose of budget day?
2 What is a 'stealth tax'? Why are stealth taxes unpopular?
3 On budget day, the government publishes an overview of the budget statement. Go to the Treasury website and download and read the executive summary of the most recent budget.
4 Budget day is usually the most important day in the year for UK fiscal policy. What is fiscal policy?

Using *AD/AS* diagrams to illustrate demand-side fiscal policy

To explain demand-side fiscal policy in more depth, it is worth reminding ourselves of the aggregate demand equation:

$$AD = C + I + G + (X - M)$$

Government spending (G) is one of the components of aggregate demand. An increase in government spending, and/or a cut in taxation, increases the size

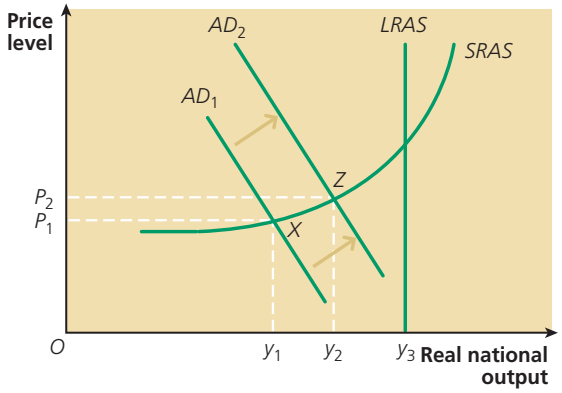


Figure 13.1 Keynesian or demand-side fiscal policy

of the budget deficit (or reduces the size of the budget surplus). Either way, an injection into the circular flow of income occurs and the effect on aggregate demand is expansionary.

Figure 13.1 illustrates the effect of such an **expansionary** (also known as reflationary) **fiscal policy**. Initially, with the aggregate demand curve in position AD_1, equilibrium national income occurs at point *X*. Real income or output is y_1, and the price level is P_1.

STUDY TIP

Expansionary and contractionary fiscal policy are changes in taxes and government spending that increase or decrease the budget deficit.

KEY TERMS

expansionary fiscal policy uses fiscal policy to increase aggregate demand and to shift the *AD* curve to the right.

contractionary fiscal policy uses fiscal policy to decrease aggregate demand and to shift the *AD* curve to the left.

To eliminate demand-deficient (also known as cyclical or Keynesian) unemployment, the government increases the budget deficit by raising the level of government spending and/or by cutting taxes. The expansionary fiscal policy shifts the *AD* curve right, from AD_1 to AD_2, and the economy moves to a new equilibrium national income at point *Z*.

SYNOPTIC LINK

The elimination of cyclical unemployment means the economy is operating on its production possibility frontier — as covered in Chapter 1.

However, the extent to which expansionary fiscal policy reflates real output (in this case, from y_1 to y_2), or creates excess demand that leads to demand-pull inflation (in this case, an increase in the price level from P_1 to P_2), depends on the shape of the *SRAS* curve, which depends on how close, initially, the economy was to its 'normal capacity' level of output (y_3 in Figure 13.1). The nearer the economy gets to its normal capacity level of output, depicted by the position of the *LRAS* curve in Figure 13.1, the greater the inflationary effect of expansionary fiscal policy and the smaller the reflationary effect. Once the normal capacity level of output is reached at y_3, a further increase in government spending or a tax cut solely inflates the price level. In this situation, real output cannot grow (except possibly temporarily) because there is no spare capacity. The economy is producing on its production possibility frontier.

SYNOPTIC LINK

Page 359 in Chapter 10 explains the meaning of 'normal capacity'. Attempts to use fiscal policy to expand the economy beyond this level may lead to inflation.

Figure 13.1 can be adapted to illustrate the effect of a deflationary or **contractionary fiscal policy**. In this case, a cut in government spending and/or an increase in taxation shifts the *AD* curve to the left. The extent to which the demand deflation results in the price level or real output falling again depends on the shape and slope of the *SRAS* curve.

TEST YOURSELF 13.2

Explain, using an *AD/AS* diagram, how a government can use fiscal policy to stimulate an economy in a recession.

EXTENSION MATERIAL

The multiplier and Keynesian fiscal policy

During the Keynesian era from the 1950s to the late 1970s, governments in many industrialised mixed economies, including the UK, based macroeconomic policy on the use of fiscal policy to manage the level of aggregate demand. This became known as **discretionary fiscal policy**.

To achieve full employment, governments deliberately ran budget deficits (setting $G > T$). This expanded aggregate demand, but sometimes too much demand 'overheated' the economy. Excess demand pulled up the price level in a demand-pull inflation, or pulled imports into the country and caused a balance of payments crisis. In these circumstances, governments were forced to reverse the thrust of fiscal policy, cutting public spending or raising taxes to reduce the level of demand in the economy. The Keynesians used demand-side fiscal policy in an isolated way (supplemented at times by monetary policy), to 'fine-tune' the level of aggregate demand in the economy. Government spending and/or taxes were changed in order to stabilise fluctuations in the economic cycle, and to try to achieve the macroeconomic objectives of full employment and economic growth, without excessive inflation or an unsustainable deterioration in the balance of payments.

The larger the multiplier, the smaller the increase in public spending or smaller the required tax cut is needed to bring about a desired increase in national income. In this case with a large multiplier, fiscal policy used as a demand management instrument can be an effective way of controlling the economy.

Unfortunately, real-world multipliers are unlikely to be much larger than a value of 1. Small multipliers mean that demand-side fiscal policy has relatively little effect on the level of aggregate demand and consequently on the economy.

A process known as **crowding out** may provide one reason for a small government spending multiplier. It is, of course, impossible to employ real resources simultaneously in both the private and the public sectors of the economy. This means that if more of the available factors of production are employed in the public sector, private-sector output must fall, assuming the economy is producing on its production possibility frontier. Employing more capital and labour in the public sector involves sacrificing the opportunity to use the same resources in private employment.

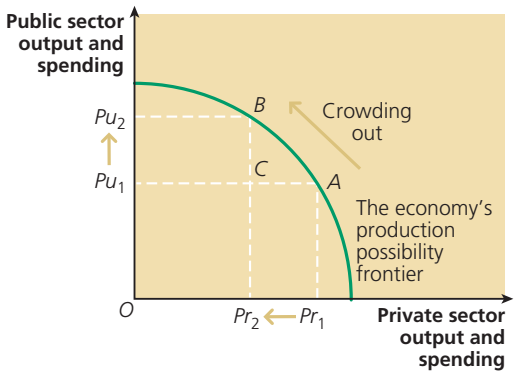


Figure 13.2 The crowding-out process

The production possibility frontier in Figure 13.2 shows maximum levels of output that can be produced with various combinations of public-sector and private-sector spending and output. Assuming there is full employment and the economy is initially at point *A*, an increase in public-sector spending from Pu_1 to Pu_2 crowds out or displaces private-sector spending, which falls from Pr_1 to Pr_2, shown at point *B*. Given these assumptions, the size of the multiplier with respect to real output is therefore zero. However, if the economy was initially producing at point *C*, inside the production possibility frontier, the multiplier is likely to be considerably larger with respect to real output. In this situation, the increase in public-sector spending from Pu_1 to Pu_2 absorbs idle or spare capacity in the economy, without reducing the resources available for the private sector to use. (Back in the 1930s, Keynes recommended fiscal expansion when the economy was in recession, or depression, but he did not advocate its use if the economy was already fully employed or operating at full capacity.)

KEY TERMS

discretionary fiscal policy involves making discrete (separate) changes to *G*, *T* and the budget deficit to manage the level of aggregate demand.

crowding out a situation in which an increase in government or public-sector spending displaces private-sector spending, with little or no increase in aggregate demand.

STUDY TIP

Most macroeconomic *PPF* diagrams show the trade-off between producing capital goods and consumer goods. Figure 13.2 is different, showing the trade-off between using scarce resources in the public sector or in the private sector.

SYNOPTIC LINK

Read about the government spending multiplier and other national income multipliers on pages 353–54 of Chapter 10. This explains that multipliers in reality are more complex than the one covered in this course.

Fiscal policy and aggregate supply

It is misleading to associate fiscal policy exclusively with Keynesian demand management. After 1979 — with the exception of the two years from 2008 to 2010 and during the Covid-19 pandemic of 2020–21, when Keynesian fiscal policy was revived to 'spend the economy out of recession' — demand-side fiscal policy gave way to **supply-side fiscal policy**.

In demand-side fiscal policy, income tax cuts stimulate aggregate demand by shifting the *AD* curve to the right. In supply-side fiscal policy, by contrast, income tax cuts increase aggregate supply via their effects on economic incentives. Supply-side fiscal policy aims to increase the economy's ability to produce and supply goods, through creating incentives to work, save, invest and be entrepreneurial. (There are other important elements to supply-side fiscal policy, including such interventionist policies as government spending on retraining schemes.) These policies will be covered in greater detail later in the chapter.

KEY TERMS

supply-side fiscal policy used to increase the economy's ability to produce and supply goods, through creating incentives to work, save, invest and be entrepreneurial. Interventionist supply-side fiscal policies, such as the financing of retraining schemes for unemployed workers, are also designed to improve supply-side performance.

supply-side policies government economic policies which aim to make markets more competitive and efficient, increase production potential and shift the *LRAS* curve to the right. Supply-side fiscal policy is arguably the most important type of supply-side policy, but there are also non-fiscal supply-side policies.

STUDY TIP

Make sure you don't confuse demand-side fiscal policy and supply-side fiscal policy.

Using an *AD/AS* diagram to illustrate supply-side fiscal policy

Along with other **supply-side policies**, which we examine in some detail later in this chapter, supply-side fiscal policy was and is used to try and shift the economy's long-run aggregate supply (*LRAS*) curve to the right, thereby increasing the economy's potential level of output. The effect of successful supply-side fiscal policy on the *LRAS* curve is shown in Figure 13.3. The economy's *LRAS* curve shifts rightwards from $LRAS_1$ to $LRAS_2$, which means that the 'normal capacity' level of output has increased (due to long-run economic growth) from y_{N1} to y_{N2}.

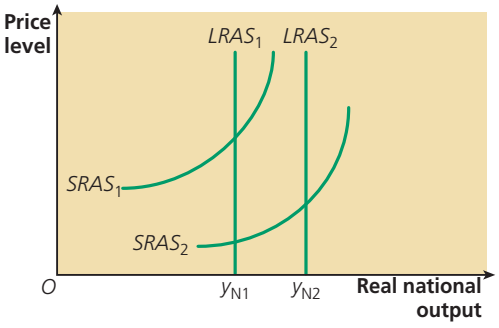


Figure 13.3 Supply-side fiscal policy shifting the *LRAS* curve to the right

Note that the diagram also shows a rightward shift of the *SRAS* curve, reflecting the improvement in productivity and falling unit costs. You could develop the diagram a stage further, by depicting the resulting fall in the price level, to show a 'good deflation', in which the price level falls because of the benign effects of successful supply-side policies and supply-side reform in the economy's private sector. However, in normal circumstances, the improvements in the supply-side performance of the economy are accompanied by rising aggregate demand, which means that the effect is to moderate inflation rather than lead to deflation.

SYNOPTIC LINK
Towards the end of the chapter we explain how supply-side fiscal policies relate to other types of supply-side policy.

STUDY TIP
Although fiscal and monetary policies can be termed either contractionary or expansionary, these terms do not apply to supply-side policies. In theory, supply-side policies are always expansionary.

Public expenditure and taxation

How government spending and taxation can affect the pattern of economic activity

Taxation raises the revenue required to finance government spending. Taxes and subsidies (which are a part of government spending) can also be used to alter the relative prices of goods and services in order to change consumption patterns and promote investment by firms in new capital goods.

Governments collect tax revenue in order to finance the provision of goods such as roads and schools which otherwise would either be under-provided or not provided at all. As we will explain shortly, a large part of government spending also transfers income and spending power between different groups in the economy: for example, from rich to poor.

SYNOPTIC LINK
Chapter 8 explained how both taxation and government spending relate to government policies to try to correct or eliminate market failures. Chapter 7 explained how taxation and government spending are linked to altering the distribution of income and wealth and reducing poverty.

Types of and reasons for public expenditure

Public expenditure has been classified into various categories. An important distinction is between investment by the government in new capital projects and infrastructure, such as new NHS hospitals and state schools, and expenditure to meet annual running costs from such projects, such as paying teachers' salaries.

A significant part of government expenditure takes the form of transfers: for example, the state pension and unemployment-related benefits. Unlike spending on capital projects and on the subsequent running costs, transfers do not involve a claim by the government on national output. Instead, spending by the government on transfers merely redistributes income and spending power from one part of the private sector to another — from taxpayers to recipients of state benefits and pensions.

A third type of public spending is interest payments on the national debt. This item of public spending rises when interest rates rise and also when the government runs a budget deficit, which is financed by new government borrowing that adds to the national debt.

SYNOPTIC LINK

The link between budget deficits and the national debt is explained later in this chapter on pages 487–88.

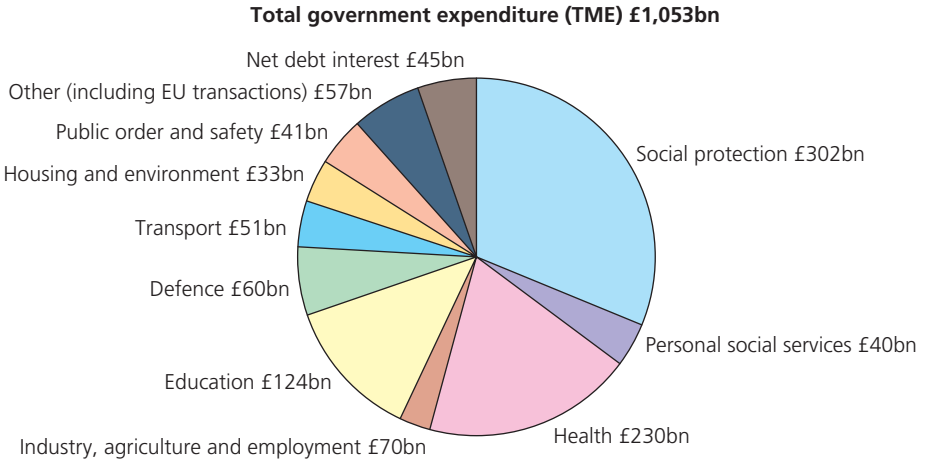
KEY TERMS

total managed expenditure the total amount that the government spends. It splits into the amount that government departments (e.g. Defence) have been allocated to spend and spending that is not controlled by a government department, including welfare, pensions and national debt interest payments.

Office for Budget Responsibility advisory public body that provides independent economic forecasts and analysis of the public finances as background to the preparation of the UK budget.

Public spending on social protection, health and education

Figure 13.4 shows estimated totals presented in the 2020 budget for different components of UK public spending, such as interest on the national debt, in the tax year 2021/22. The largest areas of public spending have been on social protection, which encompasses the financial assistance and services provided to those in need or at risk of hardship, and on health and education. In the financial year 2021/22 spending on social protection was expected to be 28.7% of central government **total managed expenditure** (TME), with spending on health and education standing respectively at 21.8% and 11.8%. The three UK government departments involved are the Departments for Work and Pensions, Health and Social Care, and Education.



Note: the figures do not add up to £1,053bn due to rounding.

Source: OBR and Treasury

Figure 13.4 Treasury estimates of government spending, 2021/22

Spending on social security, which is part of social protection spending, increased after 2007 partly because the 2008–09 recession reduced incomes and, for a time, increased the number of unemployed. Retired people now receive over half of all social security spending, with spending on the elderly forecast to continue to increase rapidly. This increase has been driven by steady growth in the number of retired people, and by the 'triple lock' which protects them from the effect of spending cuts.

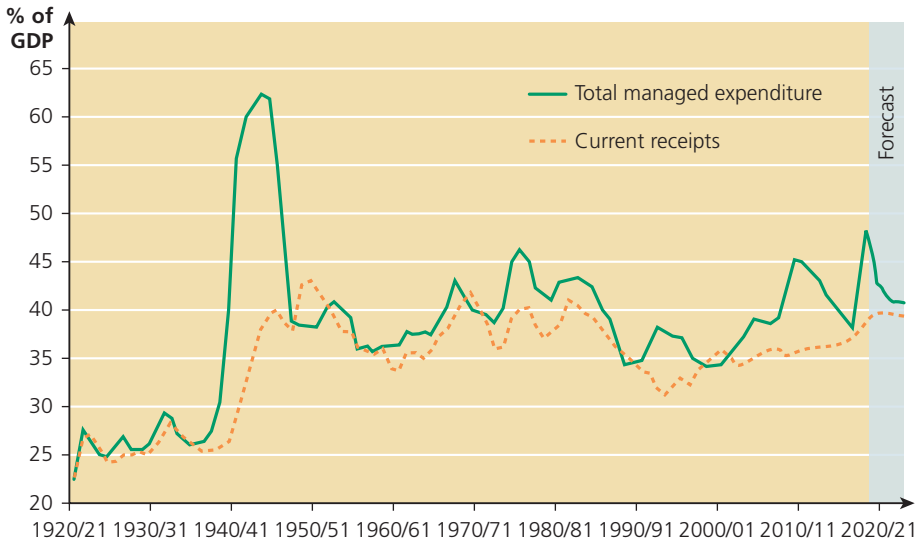
The 'triple lock' guarantees that the state pension rises by whichever is highest out of wages increases, CPI inflation, or 2.5%. Although the triple lock was temporarily suspended during the Covid-19 pandemic (as wage increases were artificially high due to workers coming off the furlough scheme), it is expected to be reinstated by 2023 at the latest. If this protection continues, the shares in total public spending of social security, the NHS and state education are likely to continue to rise. Robert Chote, the chairman of the **Office for Budget Responsibility** (OBR), has said the pensions 'triple lock' was pushing up government spending at a time when the population was ageing rapidly.

Health spending has also experienced significant growth over recent decades. Between 2002 and the recession of 2008–09, Labour governments increased spending on the NHS to around 7% of national income, but this was still below the average among the countries of the Organisation for Economic Cooperation and Development (OECD). While other departments experienced budget cuts as part of the government's austerity programme, spending on the NHS has effectively been frozen in real terms (i.e. has increased but only at the same rate as inflation) since 2010. However, despite a commitment in 2018 to increase spending on the NHS by £20 billion, increased population, longer life spans and the cost of new drugs and medical technology meant that NHS spending levels were likely to lag behind demographics-driven demand.

The forecasts and planned changes for NHS spending were overtaken by events during the Covid-19 pandemic as significant increases in spending on health were necessary in both care and the development of treatment for the effects of Covid. A rise in taxes — an increase in employee's national insurance contributions — in 2022 was meant to provide extra funding for healthcare. However, the tax was announced at a time when household budgets were under severe pressure, leading for many to call for a postponement of the tax rise. The new government in place from September 2022 reversed the decision.

As with health spending, the education budget also grew substantially under Labour governments from 2002 onwards. Like the NHS and state pensions, spending on schools has been protected, but other areas of education spending have not. Higher education has seen the largest cuts in public spending, though this has been more than offset by an increase in fees paid by students.

One reason for the changes in government spending and taxation as ratios of GDP, which are shown in Figure 13.5, lies in changes taking place in employment and unemployment, which in turn relate to the economic cycle. As Figure 13.4 showed, spending on social security (under the heading 'Social protection'), which includes unemployment-related benefits, is by far the largest single category of public spending. When the economy booms, unemployment falls, so spending on social security also falls. As a result, the cyclical component of the budget deficit becomes smaller. The reverse is true in a recession.



Sources: Bank of England, ONS, OBR

Figure 13.5 Government spending and revenue as proportions of GDP, 1920/21 to 2020/21

Demand-led spending

The pattern of public spending, and also the extent to which certain types of public spending increase while others fall, depends in part on the extent to which spending is *demand-led*. Demand-led spending — for example, on unemployment benefits — is literally led by demand, increasing when unemployment grows and falling when unemployment drops. Although the rates at which unemployment benefits and pensions are paid, and the ages at which people qualify, can be changed, demand-led spending changes according to the phases of the economic cycle and, in the case of spending on the state pension, according to how long elderly people live.

QUANTITATIVE SKILLS 13.1

Worked example: calculating government spending percentages (1)

Calculate to 2 decimal places the percentage shares of spending on social protection and education in the estimates for total government spending in 2021/22.

According to the pie chart in Figure 13.4, the estimate for spending on social protection was £302 billion out of estimated total government spending of £1,053 billion. As a percentage, this is:

$$\frac{£302bn}{£1{,}053bn} \times 100 = 28.68\%$$

The estimate for spending on education was £124 billion out of estimated total government spending of £1,053 billion. As a percentage, this is:

$$\frac{£124bn}{£1{,}053bn} \times 100 = 11.78\%$$

QUANTITATIVE SKILLS 13.2

Worked example: calculating government spending percentages (2)

In 2021/22, the government expected to spend £1,053 billion. Nominal gross domestic product (GDP) was expected to be £2,317.05 billion.

Calculate, to 2 decimal places, total planned government expenditure as a percentage of nominal GDP.

The correct answer is:

$$\frac{\text{total planned expenditure}}{\text{nominal GDP}} \times 100$$

which is:

$$\frac{£1{,}053bn}{£2{,}317.05bn} \times 100 = 45.45\%$$

Transfer payments made by the government

A large part of government expenditure takes the form of transfer payments: for example, the state pension and unemployment-related benefits. As the

name indicates, in this context (though not in the way the term 'transfers' is used in the current account of the balance of payments), transfer payments are a redistribution of spending power from taxpayers in general to those receiving welfare benefits. By contrast, government spending on new hospitals or schools directly increases national output. As a generalisation, income taxes and transfers reduce the disposable incomes of those in work and increase the disposable incomes and spending power of those living on welfare benefits, who are very often out of work.

Transfers are the major part of the item 'social protection', which at over £300 billion in 2021/22 dwarfed all other types of public spending, including health and education. When transfers, including debt interest, are excluded, government spending falls as a proportion of GDP. This figure is a more accurate measure of the share of national output directly commanded by the state (and thus unavailable for use in the private sector) to produce the hospitals, roads and other goods and services which government collectively provides and finances, for the most part, out of taxation.

Debt interest

Debt interest is made up of payments by the government to people who have lent to the state (i.e. to holders of the national debt). In 2021/22, interest payments on the national debt were expected to be £45 billion, or just over 4% of public spending. This item of public spending, which rises when interest rates rise, and falls when interest rates are cut, is a transfer from taxpayers in general to people who lend their savings to the government. Total interest payments are affected by the general level of interest rates, which are heavily influenced by the level of Bank Rate set by monetary policy. In terms of fiscal policy, if the national debt (relative to nominal GDP) can be reduced, debt interest as a fraction of nominal GDP also falls — provided interest rates don't rise. Conversely, if the national debt rises faster than nominal GDP, debt interest rises as a fraction of real GDP — provided interest rates don't fall.

How taxation affects the pattern of economic activity

Figure 13.6 shows the total amount of revenue the UK government expected to collect from different taxes in 2019/20. By studying the pie graph, you can calculate the relative or proportionate importance of each tax. The three main categories of tax are taxes on income, taxes on spending or expenditure, and taxes on capital.

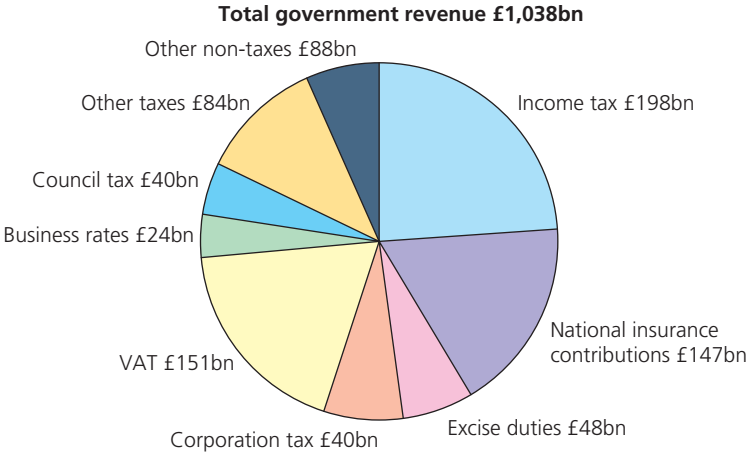


Figure 13.6 Treasury estimates of tax revenue, 2021/22

Taxes on income include not only personal income tax (income tax in Figure 13.6) but also national insurance contributions and corporation tax. National insurance contributions (which in legal terms are not strictly a tax) are a second type of personal income tax, which may eventually be merged into personal income tax. Corporation tax is a tax paid by companies on their profits.

The two main indirect taxes are value added tax (VAT) and excise duties on goods such as motor fuels, alcohol drinks and tobacco. In 2021/22 these were expected to raise £151 billion and £48 billion in tax revenues. These figures are probably lower than would normally be expected. This is the result of prolonged depressed spending on items during the Covid-19 pandemic of 2020–21 and also the temporary cuts in some of these taxes designed to support businesses during this period.

Taxes on wealth and capital are not very important in the UK. Council tax, which is a tax on property, and business rates are the only wealth taxes shown in Figure 13.6. Inheritance tax, which is a tax on wealth given from the dead to the living, is included in the category 'Other taxes' in the pie chart. Much of income tax revenue is, of course, paid when people work for a living, producing useful goods and services. Wealth taxes, which by contrast stem from the ownership of assets such as houses and stocks and shares, reflect the accumulation or build-up of personal, household and business wealth in the economy.

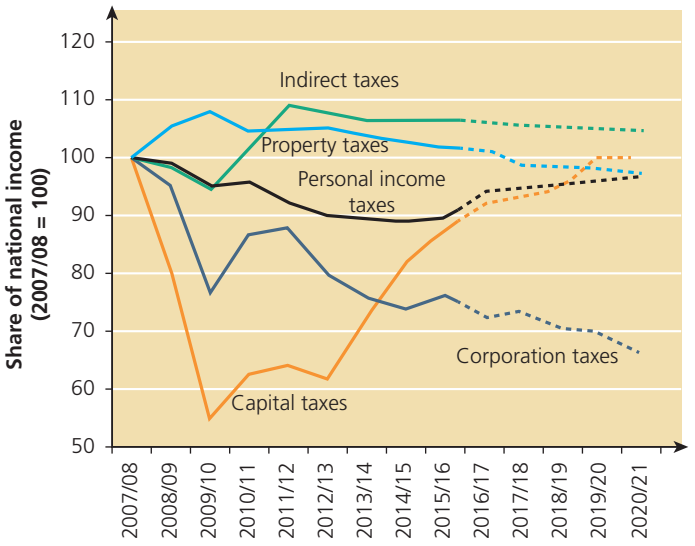
Types of and reasons for taxation

A tax is a compulsory levy made by a government to pay for its activities. Taxes are therefore used to finance the different types of public expenditure we have just described.

Taxation is the principal source of government revenue for most economies. Other forms of income would come from the sale of services offered by the government (such as sales of medical prescriptions) or interest or dividends received from UK government investments.

Figure 13.7, which has been taken from an Institute of Fiscal Studies (IFS) briefing note, 'The changing composition of UK tax revenues', shows how the composition of different categories of tax receipts has changed, or is expected to change, between 2007/08 and 2020/21. (Note: at the time of publication, no current data are available for the figures forecasted for 2020/21.)

Figure 13.7 Composition of UK government tax receipts, 2007/08 to 2020/21 (forecast)



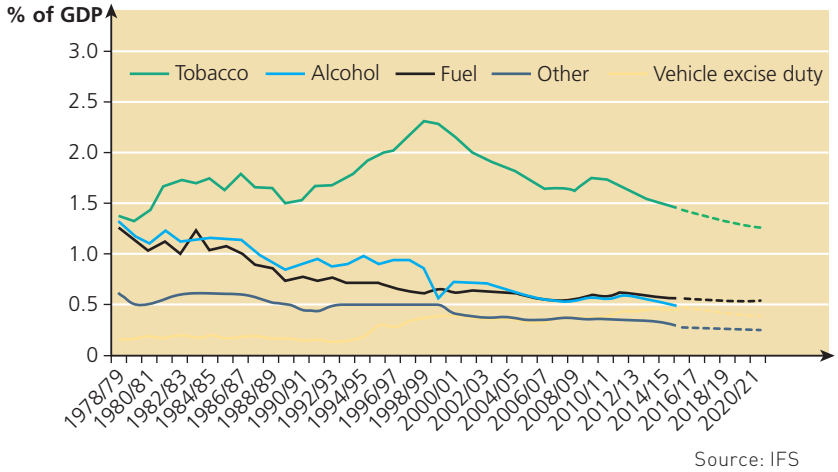
Source: IFS

TEST YOURSELF 13.3

Look carefully at Figure 13.7.

What can you conclude from the way the data are presented, and especially from the fact that the index numbers for all the types of tax equal 100 in the financial year 2007/08?

Figure 13.8, taken from the *IFS Green Budget: February 2016*, shows revenues from excise duties as a percentage of national income from 1978/79 to 2020/21. In 2018/19 the UK government raised £49 billion from excise duties, but as can be seen in Figure 13.8 the long-term revenue received by the Treasury from these duties is in decline. In the future, with people choosing to live healthier lives and moving away from combustion engines and fossil fuels, the government is going to have to find new sources of taxation to finance public expenditure. This will become increasingly important as the sale of petrol- and diesel-powered cars will be phased out by 2030 in the UK.



Source: IFS

Figure 13.8 UK tax revenue from excise duties, 1978–2014 to 2020/21 (forecast)

We have previously given reasons for governments choosing to undertake different types of public spending, so these are also reasons for taxation. To repeat, governments tax in order to achieve economic goals such as altering the distributions of income and wealth and the management of aggregate demand. We will now look at direct and indirect taxation; progressive, proportionate and regressive taxes; and the principles of taxation.

EXTENSION MATERIAL

More detail on the structure of UK taxation

Some of the main taxes levied in the UK are described below.

Income tax

This is the UK's main direct tax, which is paid on earnings, pensions, benefits, savings and investment income, and rents. For most people, the first slice of income is untaxed. In the tax year 2022/23 in England, Wales and Northern Ireland, the personal tax allowance (untaxed income) for people with an income of £100,000 or less was £12,570, but for higher-income earners the personal tax allowance tapered away, dropping by £1 for every £2 earned over £100,000. In the same year, the standard rate of income tax (20%) was levied on individual income above the personal tax allowance but below £50,000. Higher rates of tax were levied on personal incomes above £50,270, namely 40% on the slice of income

between £50,271 and £150,000, and 45% on income above £150,000.

National insurance

National insurance contributions (NICs), first started in 1911, were initially introduced to finance a government safety net for workers who fell on hard times. Anyone needing cash for medical treatment, or because they had lost their job, could claim from the fund. Over the years, the system has changed. The national insurance fund is now used to help finance the NHS and welfare benefits, including unemployment benefit and the state pension. However, NICs are insufficient for this purpose and have to be supplemented from general taxation.

In strict legal terms, NICs are not a tax. Consequently, on a number of occasions UK governments have increased NICs, while leaving personal income tax rates and allowances unchanged, and then claimed that they have not increased income tax! Most people, however, regard NICs as a supplementary form of income tax. Over the years, NICs have generally been regressive, and this has reduced the progressivity of the direct tax system, provided we regard NICs as a form of tax. The terms 'regressive' and 'progressive' are covered later in this section.

In the future, income tax and national insurance could be merged by the Conservative government under plans to simplify the tax system. If this happens, it will be the biggest shake-up in the UK tax system certainly since the introduction of VAT in 1973, when it replaced an older percentage sales tax called purchase tax.

Corporation tax

This is a tax on company profits. It is levied at a lower rate than personal income tax. In recent years corporation tax rates have been cut. At present the main rate of corporation tax is 19% (a cut was planned for 2020 but this was cancelled to help raise more tax revenue).

Value added tax (VAT)

Unlike income tax and NICs, which are direct taxes, VAT is an indirect tax, currently paid at a standard percentage rate of 20% on spending on most goods and services. VAT is set at a lower rate of 5% on gas and electricity. It is not currently levied on books, children's clothes and most food and drink (though some items of food and drink are subject to the 20% rate, such as alcoholic drinks, crisps and savoury snacks).

Excise duties

These are another form of expenditure tax (and indirect tax) paid on tobacco, alcohol, petrol and diesel fuel. Unlike VAT, which is an *ad valorem* percentage tax, these so-called 'sin taxes' are specific taxes or unit taxes levied on physical quantities of a good. For example, in 2022 an excise duty of £297.57 was levied on each hectolitre (100 litres) of wine with an alcohol content between 5% and 15%. With motor fuels, the excise duty levied on petrol is lower than that on diesel fuel. As inflation erodes the real value of excise duties, the government usually increases them in the annual budget. Excise duties are generally regressive because they take a larger percentage of the income of the low-paid.

Stamp duty

Stamp duty land tax (SDLT) is paid on the purchase of property. The 2022 SDLT threshold was £125,000 for residential properties and £150,000 for non-residential land and properties. Higher rates of SDLT are used for higher-value properties, second homes and buy-to-let properties, and exemptions are made for first-time buyers (as long as the price of the property is below a certain level).

Inheritance tax

When a person dies, inheritance tax, which is a tax on wealth, is paid on their estate at 40% on any value above £325,000 during 2022/23.

Capital gains tax

This is levied on the profit a person makes when selling possessions or investments. The tax is paid on the increase in value from the date the assets were acquired to the date when they are sold or given away. Some forms of wealth are exempt from capital gains tax — for example, a main home, or if the gain is less than £12,300 during the tax year.

Council tax

This is a tax charged by local authorities on the value of houses and flats to the people who live in the properties. Council tax rates are 'banded': the lowest rate is charged on Band A properties with a value of up to £40,000 (in 1993 when the tax was introduced — though later in Wales), while the highest rate is paid on Band H properties with a value above £320,000 in 1993. Since they have never been revised, these bands are now hopelessly out of date. As house prices have increased enormously since 1993, the values for each band are largely irrelevant. If houses have been significantly altered/extended since that date, the band they are in could well be 'incorrect' for the relative value of the property.

KEY TERMS

direct tax a tax that cannot be shifted by the person legally liable to pay the tax onto someone else. Direct taxes are levied on income and wealth.

indirect tax a tax that can be shifted by the person legally liable to pay the tax onto someone else (e.g. through raising the price of a good being sold by the taxpayer). Indirect taxes are levied on spending.

regressive taxation when the proportion of income paid in tax falls as income increases.

Direct and indirect taxation

Income tax is an example of a **direct tax** because the person who receives and benefits from the income is liable in law to have to pay the tax to the government; they cannot pass the tax on to someone else. Corporation tax and national insurance contributions are other examples of direct taxes.

By contrast, *most* taxes on spending, such as value added tax (VAT) and excise duties, are **indirect taxes**. This is because the seller of the good, and not the buyer who benefits from its consumption, is liable to pay the tax.

SYNOPTIC LINK

As discussed in Chapter 3, firms try to raise the prices they charge customers in order to recoup the tax revenue they pay to the government. When this happens, the buyers of the good indirectly pay some or all of the tax, via the higher prices the sellers now charge.

Progressive, regressive and proportionate taxation

If the government decides that the distributions of income and wealth produced by free-market forces are undesirable, taxation and transfers in its public spending programme can be used to modify these distributions and to try to reduce this market failure resulting from 'inequity'.

Progressive taxation

Until quite recently, all UK governments used progressive taxation and a policy of transfers of income to the less well-off, in a deliberate attempt — albeit with limited success — to reduce inequalities in the distribution of income. A tax is progressive when the proportion of income paid in tax rises as income increases (see Figure 13.9). Progressive taxation, combined with transfers to lower-income groups, reduces the spending power of the rich, while increasing that of the poor. However, some taxes, particularly those designed to reduce consumption of the demerit goods alcohol and tobacco, are regressive and fall more heavily on the poor.

Regressive taxation

Cigarette duty is an example of **regressive taxation**. The low-paid lose a greater proportion of their incomes in tax than rich people when buying cigarettes and alcohol. VAT is another example of a regressive tax.

To understand why the tax on cigarettes is regressive, compare two smokers, one with an income of £100 a week and the other with an income of £1,000 a week. Both smoke 20 cigarettes a week, which in 2018 cost £8.50, of which £6.98 was tax — both excise duty and VAT. This is 6.98% of the poor smoker's income, compared to 0.698% of the rich smoker's income. The regressive nature of taxes such as these partially offsets the fact that income tax is generally progressive. Tax on cigarettes is usually increased faster than the rate of inflation, which means the regressive nature of tobacco duty is likely to increase over time. Many see this as not a problem based on the nature of tobacco as a demerit good. Currently the tax on a pack of 20 cigarettes is 16.5% of the retail price plus an additional £5.33. An average pack costs around £12, which means the majority of the retail price is tax.



Cigarettes are subject to regressive taxation

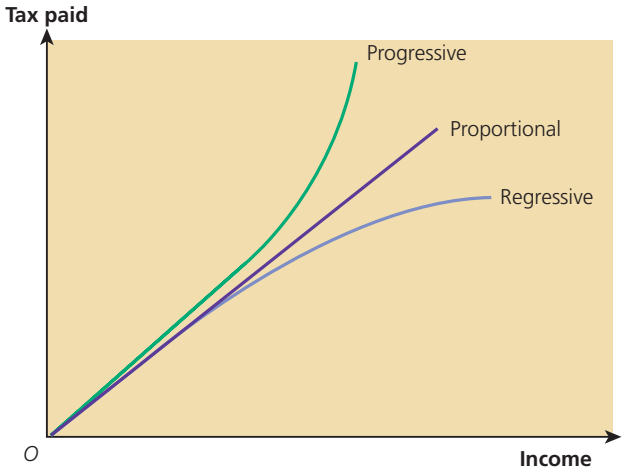


Figure 13.9 Progressive, proportionate and regressive taxation

Proportional taxation

In recent years, many economists and politicians, have advocated the introduction of **proportional taxation**, which in the case of a proportional income tax is sometimes called a 'flat tax' (e.g. income tax at the 'flat rate' of 10%). A proportional income tax system significantly benefits high-income groups over the rest of society. The equity of a system that allows millionaires to pay the same rate of income tax as ordinary workers is questionable. Proportional taxation would result in governments facing significant budget deficits, which in the long term would mean public expenditure needed to shrink.

Relating progressive, regressive and proportional taxation to marginal and average tax rates

For particular taxes such as income tax or inheritance tax, we can identify whether the tax is progressive, regressive or proportional by examining the relationship between the *average* rate at which the tax is levied and the *marginal rate*. In a progressive income tax system, the marginal tax rate is higher than the average tax rate, which means that the average rate, which measures the proportion of income paid in tax, rises as income increases. Conversely, in a regressive income tax system, the marginal rate of tax is less than the average rate, while the two are equal in the case of a proportionate tax.

For income tax, the average tax rate is calculated as total tax paid divided by total income. By contrast, the marginal tax rate is the tax paid on the last pound of income earned, calculated as the change in tax paid divided by the change in income.

$$\text{average tax rate} = \frac{\text{tax paid}}{\text{income}} \text{ or } \frac{T}{Y}$$

$$\text{marginal tax rate} = \frac{\Delta\text{tax paid}}{\Delta\text{income}} \text{ or } \frac{\Delta T}{\Delta Y}$$

As a general rule, the average tax rate indicates the overall burden of the tax on the taxpayer, but the marginal rate may significantly affect economic choices and decision making. In the case of income tax, the marginal tax rate influences the choice between work and leisure when deciding how much labour to supply. It also influences decisions on whether to spend income on consumption or to save.

KEY TERM
proportional taxation when the proportion of income paid in tax stays the same as income increases.

STUDY TIP
Make sure you can explain how progressive taxation and transfers have affected the economy and can evaluate whether the effects have been good or bad.

QUANTITATIVE SKILLS 13.3

Worked example: calculating the marginal tax rate

Assuming that all income between £30,000 and £38,000 is taxed at the same rate, if total tax paid is £4,000 when total income is £30,000 and £6,000 when income is £38,000, what is the marginal tax rate?

Income rises by £8,000 and the tax paid rises by £2,000. The marginal tax rate is:

$$\frac{\Delta T}{\Delta Y} \text{ or } \frac{£2{,}000}{£8{,}000}$$

which is 25%.

The principles of taxation

KEY TERM
principle of taxation (also known as a **canon of taxation**) a criterion used for judging whether a tax is good or bad.

Taxpayers commonly view all taxes as 'bad', in the sense that they do not enjoy paying them, although most realise that taxation is necessary in order to provide for the useful goods and services provided by the government. A starting point for analysing and evaluating whether a tax is 'good' or 'bad' is Adam Smith's four **principles of taxation**, which are also known as the **canons of taxation**. Adam Smith suggested that taxation should be equitable, economical, convenient and certain, and to these we may add the principles of efficiency and flexibility. A 'good' tax meets as many of these principles as possible, although because of conflicts and trade-offs, it is usually impossible for a tax to meet them all at the same time. A 'bad' tax meets few if any of the guiding principles of taxation.

SYNOPTIC LINK
Adam Smith appeared in earlier chapters (2, 4 and 5) when looking at specialisation, the invisible hand and competition.

- **Economy** means a tax should be cheap to collect in relation to the revenue it yields.
- **Convenience** and **certainty** mean that a tax should be convenient for taxpayers to pay and that taxpayers should be reasonably sure of the amount of tax they will be required to pay to the government.
- **Equity** means a tax system should be fair, although there may be different and possibly conflicting interpretations of what is fair or equitable. Specifically, a particular tax should be based on the taxpayer's *ability to pay*. This principle is one of the justifications of progressive taxation, since the rich have a greater ability to pay than the poor.
- **Efficiency** requires a tax to achieve its desired objective(s) with minimum undesired side-effects or unintended consequences. The disincentive effect on effort can be thought of as an unintended consequence of high rates of income tax.
- Finally, to comply with the principle of **flexibility**, a tax must be easy to change to meet new circumstances.

TEST YOURSELF 13.4
Distinguish between the principles of taxation and the aims or objectives of taxation.

The role and merits of different UK taxes

A large proportion of people's income goes in paying taxes. Tax accounts for around a third of the money people earn. Direct taxes such as income tax and

national insurance account for 20%. The rest goes mainly on VAT, duty on alcohol and petrol, council tax and other indirect taxes.

Taxation is the principal source of government revenue for most economies. In the UK about 93% of total taxation is levied by central government, with local government taxation (currently the council tax and business rates) accounting for the remaining 7%. In the financial year 2021/22, all but £88 billion of the total expected revenue of £1,038 billion was expected to come from taxation. The non-tax income was expected to come from other sources, such as interest and dividends which the government receives. UK governments have also received revenue from privatising state-owned firms and industries. However, in order to keep down the figure for public spending, UK governments perform the accounting trick of classifying these receipts as 'negative public expenditure' rather than as a source of government revenue.

We will now look in turn at taxes on income, taxes on spending and taxes on capital and wealth.

Taxes on income

If you refer back to Figure 13.6, the chart shows that taxes on income, which are the most important category of tax in the UK, include corporation tax and national insurance contributions as well as personal income tax. These are all direct taxes. This means that the person or organisation being paid the income is directly liable to pay the tax to the government. Failure to declare income is tax *evasion*, which is illegal.

With regard to some of the principles of taxation outlined earlier, for most wage and salary earners, income tax is cheap to collect, it is convenient and certain for the taxpayer, and, if progressive, it is equitable in the sense that it reflects taxpayers' ability to pay. In addition, when the basic tax threshold is set at a relatively high level, people who receive very low incomes can be taken out of the 'tax net', thereby paying zero income tax. For most wage and salary earners, income tax is collected by the government through 'pay as you earn' (PAYE). This makes personal income tax cheap to collect, as employers bear most of the costs.

Nevertheless, for some in society, income tax has been relatively easy to avoid and evade. Tax avoidance is not illegal but may allow some taxpayers to exploit unintended loopholes in the tax system. People on relatively low incomes may work 'strictly for cash' (or 'cash in hand'), where physical money is exchanged for work carried out. This means that there is no record of the money being paid into a bank account and no receipt for the work is issued. If sub-standard work is carried out, the customer cannot 'prove' that the person did the work. Higher earners may evade tax through signing up to tax-efficient schemes provided by their financial advisers.

The opportunities for avoiding and/or evading paying tax are, of course, among the disadvantages of income tax. Additionally, a highly progressive income tax may lead to undesirable 'unintended consequences' — for example, the disincentivising of hard work, risk taking and entrepreneurial effort. We return to this in the second half of the chapter, in our discussion of supply-side economic policies.

However, as Figure 13.6 clearly shows, personal income tax has long been the main source of government tax revenue in the UK. In part, this reflects the fact that for personal income tax, the 'tax base' is extremely wide in the sense that millions of people receive income that can be taxed. Total government revenue

from the taxation of income is actually much higher than the revenue collected from personal income tax. Companies pay corporation tax on their income or profits and most of their employees pay national insurance contributions (NICs) as well as personal income tax to the government.

QUANTITATIVE SKILLS 13.4

Worked example: calculating the mean and median

Table 13.1 shows the weekly incomes of the 20 members of an economics class at an adult education college.

Table 13.1 Weekly incomes of an economics class

Class member	Weekly income (£)	Class member	Weekly income (£)
Student 1	220	Student 11	500
Student 2	1,000	Student 12	370
Student 3	800	Student 13	80
Student 4	340	Student 14	1,200
Student 5	620	Student 15	50
Student 6	800	Student 16	570
Student 7	400	Student 17	920
Student 8	680	Student 18	30
Student 9	730	Student 19	250
Student 10	690	Student 20	820

Calculate the mean and median weekly incomes for the class.

As we first explained in Chapter 3, the mean and the median are two kinds of 'average'. The mean class income can be calculated by adding up the incomes of all the members of the class (in this case £11,070) and dividing this number by 20 (the total number of class members). The mean value for all the members of the economics class is £553.50.

The 'median' is the 'middle' value in the list of numbers. To find the median, you should list the numbers in numerical order (£30; £50; £80; £220; £250; £340; £370; £400; £500; £570; £620; £680; £690; £730; £800; £800; £820; £920; £1,000 and £1,200), and then locate the middle value. In this example, because there is an even number of students in the class (20), there are two middle values, £570 and £620. The median is midway between the two: £595.

TEST YOURSELF 13.5

Chloé wants to achieve a mean score of at least 90% on her six economics multiple-choice tests. Her first five multiple-choice test scores are: 95%, 88%, 97%, 82%, 91%. What is the minimum score that Chloé needs to get on her last multiple-choice test in order to reach her target?

Taxes on spending

As Figure 13.6 again shows, after income taxes, VAT and 'excise duties' make up the second most important group of taxes in the UK. VAT is an *ad valorem tax*, while excise duties, such as the duty charged on spending on tobacco, are almost always *specific* or *unit taxes*, levied not on a good's price, but on the physical quantity of the good.

An advantage of taxes on spending, especially excise duties, is that they can be used to encourage people to switch their spending away from goods deemed to be undesirable to those regarded as 'good' for consumers — for example, away from demerit goods and towards merit goods. Some economists, mainly of a free-market persuasion, believe this to be a disadvantage rather than an advantage. They dislike governments claiming to know better than ordinary individuals what is good for those whom they govern. VAT, by contrast, is a generally neutral tax, in that it is levied on spending on most goods and services, with the result that a change in the VAT rate has little effect on patterns of expenditure.

Many economists believe that, unless there is a good reason to affect people's choices through the tax system, the taxation of goods should conform to the principle of neutrality. Reasons for taxing some products at a higher rate than others could relate to demerit goods and externalities. Another reason could be to influence the distribution of income, imposing zero tax on basic necessities and high taxes on luxury goods.

In any case, using indirect taxes to encourage people to alter their expenditure patterns becomes less effective if more and more goods are taxed at the same rate. Although this widens the 'tax base', which allows the government to collect more tax revenue, the tax becomes less useful as a tool for influencing the pattern of consumer spending.

Indirect taxes can also be evaded: for example, when a builder says they will not include VAT in the price they are quoting to a householder, as long as the householder pays in cash. However, it is generally less easy to evade and avoid indirect taxes than direct taxes.

QUANTITATIVE SKILLS 13.5

Worked example: calculating percentage change

Table 13.2 shows the changes in a government's total revenue from indirect taxes, expressed as index numbers, for the period from 2018 to 2023.

Table 13.2 Index of revenue from taxation, 2018–23

Year	Revenue from indirect taxation (2017 = 100)
2018	96
2019	94
2020	100
2021	102
2022	106
2023	104

Calculate to 2 decimal places the percentage change in revenue from indirect taxation between (a) 2018 and 2023, (b) 2020 and 2023.

Students often confuse percentage changes with changes in index points, possibly because percentages add up to 100%, while the base index number is normally 100. (There are exceptions — the base index number in the Financial Times/Stock Exchange index of the 100 leading shares (the FTSE 100 or 'Footsie 100' index) is 1,000.)

(a) Between 2018 and 2023 the index point values changed from 96 to 104, i.e. by 8 index points. However, this is *not* an 8% change. The percentage change is the change in index points divided by the initial index number, which is then multiplied by 100. In this case it is (8/96) × 100, which is approximately 8.33%.

(b) In this case, because we are comparing 104 with the base year index number of 100, the change in index points is 4 and the percentage change is 4%.

Taxes on capital and wealth

Taxes on capital and wealth are also direct taxes. Some taxes on capital and wealth can be avoided and evaded, but with others it is less easy to do so. Wealth in the form of cash can be 'money laundered' when given from one person to another. This means that the origin of money obtained illegally or without paying tax is concealed by converting it to a legitimate source. It is then added into the normal financial system, where it no longer appears as suspicious. On the other hand, wealth in the form of property is less easily hidden. However, taxing property would work best if the government had up-to-date values for properties, which is not easy given how quickly their values change. Successive UK governments have shown little interest in updating property values. The last time houses were valued for the purpose of collecting council tax, which is the main UK tax on property, was back in 1991 (though later for Wales), since when house prices have risen rapidly in most UK regions, especially in London and the southeast.

Some general comments on the structure of UK taxation

The financial crisis and the 2008–09 recession saw the government's tax revenues collapse and the budget deficit increase sharply. Tax receipts from capital taxes (which include stamp duties, capital gains tax and inheritance tax) declined rapidly and have taken over a decade to recover. Corporation tax revenues fell significantly during the recession and appear to be in long-term decline. Tax revenue from personal income (mostly income tax and national insurance) has struggled to recover from the crisis and is increasingly dependent on higher-income groups.

The same happened again during the pandemic of 2020–21. Large increases in government spending on support for individuals and businesses (such as the furlough scheme) were combined with rapid falls in taxation as spending on goods and services was restricted and remained low in the recovery period. However, the recovery in the government's budget balance occurred more rapidly from the 2021/22 tax year onwards, mainly due to prices rising quickly, which meant that VAT receipts rose faster than expected.

Taxes on spending are now the main source of government tax revenue. VAT was 8% in the late 1970s but has gradually risen, and since 2012 has been 20%. Likewise, property taxes (council taxes and business rates) have increased as a source of government revenue. Overall, regressive taxes, which affect the disposable income of low-income groups, have increased since the recession.

SYNOPTIC LINK

For a reminder of how taxes (and their opposite, subsidies) can affect resource allocation, refer back to the coverage of merit goods, demerit goods, public goods and externalities in Chapter 8. Consider also how you might apply the concept of allocative efficiency (equating a good's price with its marginal social cost of production) in your analysis of how taxes and subsidies can be used to correct market failure by improving resource allocation.

The role and merits of taxes in affecting the distribution of income and wealth

With regard to the distributions of income and wealth, in past decades UK governments used taxation and public spending to try to reduce inequality between rich and poor. More recently, although governments have been less concerned about reducing inequality, the combined effects of taxes and public spending have made the distribution of income much more equal than if governments did not intervene. Under the influence of supply-side thinking, government policy has been affected by the conflict between two of the principles of taxation mentioned previously: *efficiency* and *equity*.

Efficiency requires greater incentives for work and enterprise in order to increase the UK's growth rate. However, when progressive taxation and transfers to people who are less wealthy are used to try to make the distribution of income more equitable, it can mean that people have less incentive to work hard and to take risks by engaging in entrepreneurial activity. Moreover, the ease with which the poor can claim welfare benefits and the level at which they are available can create a situation in which the poor choose unemployment and state benefits in preference to work and low wages. This is not necessarily due to people being lazy but may be a rational choice given the loss in benefits if a job is obtained, and the extra costs, such as transport to and from work, and extra childcare costs that will be incurred once someone is working.

The supply-side view

Many supply-side economists and politicians have argued that income tax rates and benefit rates should both be reduced. They believe that tax and benefit cuts would alter the work/leisure choice in favour of supplying labour, particularly for benefit claimants who lack the skills necessary for high-paid jobs. They believe that to make everyone eventually better off, it is necessary to increase the gap between the amount people earn when they are in work and what they receive when they are out of work. This means making those who are poor, because they are unemployed, worse off.

According to this view, increased inequality is necessary to create incentives to facilitate economic growth, from which all will eventually benefit. Through a 'trickledown' effect, the poor eventually end up better off in absolute terms but, because inequalities have widened, they are still relatively worse off compared to the rich.

According to data from the Office for National Statistics (ONS), inequality increased significantly in the 1980s and stabilised in the late 1990s. This was the intended outcome of government supply-side fiscal policy. In the first two decades of the twenty-first century inequality has remained high, although it has narrowed in the years following the financial crisis because the incomes of high-income groups have fallen faster than those of low-income groups. Nonetheless, the UK remains one of the most unequal countries in Europe in terms of disposable income. According to the OECD, in 2019 the UK's Gini coefficient was 37% compared to France and Germany's 29% and Italy's 33%.

When considering the merits and demerits of different types of taxation, a starting point is to consider the extent to which a particular tax satisfies the various principles of taxation we have briefly explained. As we mentioned earlier, a 'good' tax meets as many of these principles as possible, whereas a 'bad' tax meets few, if any, of the guiding principles.

TEST YOURSELF 13.6
Explain how equality and equity will be affected if a government cuts income tax rates and increases indirect taxes such as VAT.

SYNOPTIC LINK
Income inequality is discussed in the context of the Lorenz curve and the Gini coefficient in Chapter 7.

STUDY TIP
Practise drawing the Lorenz curve to illustrate how pro-market supply-side policies have resulted in higher levels of income inequality in the UK. Remember always to integrate your diagrams into your written answers.

TEST YOURSELF 13.7
Using Figure 13.6, calculate the percentage shares of income tax and corporation tax in estimated total government revenue in 2021/22.

The budget deficit and the national debt

Revisiting the budget deficit

Towards the beginning of the chapter we took a first look at the budget deficit, together with budget surpluses and balanced budgets. We defined a budget deficit as the difference between government spending and revenue from taxes and other revenue sources when government spending exceeds revenue ($G > T$). In this section, we relate budget deficits to the national debt, before in the following sections taking a look at some other aspects of budget deficits.

The national debt

KEY TERM
national debt the *stock* of all past government borrowing that has not been paid back.

The **national debt** is a rather misleading term. It is *not* the debt of the whole nation, or even of the whole of the public sector (the public-sector debt), with which it is often confused. Rather, the national debt is the total stock of central government debt — the accumulated stock of central government borrowing which has built up over the years and which the government has not yet paid back.

In the Keynesian era, when almost continuous deficit financing led to a steadily rising nominal national debt in the UK, the national debt was not regarded as a burden on the population and reduction of the national debt was not really a fiscal policy objective. However, during this period the national debt as a proportion of nominal GDP was falling because the economy was growing and experiencing inflation. By contrast, in more recent years — and related to the aim of trying to reduce the levels of both public spending and public-sector borrowing as proportions of national output — national debt reduction has become an important part of supply-side fiscal policy. Free-market supply-side economists are in favour of low levels of government spending and low levels of taxation, and are also, usually, in favour of the government running a balanced budget.

The relationship between the budget balance and the national debt

In recent decades, UK governments have usually run budget deficits. In this situation, the *flow* of public-sector borrowing which finances the budget deficit builds up a *stock* of accumulated debt. The central government's accumulated

debt is the national debt. Government ministers as high as the prime minister — the formal head of the Treasury — sometimes confuse the two 'd' words: deficit and debt.

STUDY TIP

Make sure that you understand the difference between a flow and a stock, and that you are then able to provide other examples of economic flows and stocks.

It is useful also to distinguish between the *nominal* or *money* values of the budget deficit and the national debt as a percentage of nominal GDP — the 'debt to GDP' ratio. The debt to GDP ratio is an indicator of the burden of the national debt on the economy. With an understanding of these differences, you can appreciate that while the nominal national debt may be rising, the national debt as a percentage of nominal GDP may be falling. This happens when nominal GDP rises faster than the nominal debt, either because of economic growth (which is good) or inflation (which may be bad), or both.

In the depths of recession before 2010, both the budget deficit and the national debt (and the wider public-sector debt) were growing. More recently, the deficit fell by a third (a fact much trumpeted by the Treasury), but the national debt continued to grow. Between 2015 and 2020 the national debt was approximately 80% of GDP. However, given the large deficits of 2020 and 2021 to fund the Covid-19 pandemic support, the national debt has risen further and in mid-2022 was around 100% of GDP, and set to rise further over the next few years. To understand the reason for this, you must first understand the difference between stocks and flows. A budget deficit is an example of an economic flow. This means that even when the deficit is falling, provided it is still positive (i.e. not a budget surplus), the flow of new borrowing that finances the budget deficit adds to the stock of the national debt.

STUDY TIP

Although a national debt to GDP ratio of 100% may sound worrying, think about the fact that many households in the UK have mortgages well in excess of their annual earnings: 100% may be high but it is perhaps not as catastrophic as you might initially think given that the debt is likely to be paid back over a number of years. The national debt does, however, need to be serviced (interest paid on the debt), which is a burden on the economy.

As it increases the national debt, a budget deficit generally increases the total interest payments the government has to pay to savers who have lent to the government. Conversely, a budget surplus allows the government to reduce the national debt by paying back a fraction of past borrowing. As mentioned in the previous section, changes in interest rates will also affect the outcome.

TEST YOURSELF 13.8

Evaluate the view that budget deficits and government borrowing are a cause for concern.

QUANTITATIVE SKILLS 13.6

Worked example: interpreting national debt statistics

In an economy, the cash value of the national debt increased from £1,000 billion in 2022 to £1,200 billion in 2023, while the nominal value of the country's GDP increased by 10% over the same period from its £1,600 billion value in 2022.

(a) What was the nominal national debt as a ratio of nominal GDP in 2022?
(b) What happened to this ratio over the year to 2023?
(c) What can we conclude about any change in the burden of the national debt over the year in question?

(a) In 2022 the ratio was £1,000bn/£1,600bn, which is 0.625 or 62.5%.
(b) During 2023, the ratio increased, calculated from £1,200bn/£1,760bn, which is 0.6818 or 68.18%.
(c) The burden of the national debt, measured by the nominal (or cash) value of the debt as a ratio of nominal national income, increased from 62.5% to 68.18%.

Cyclical and structural budget deficits and surpluses

KEY TERMS

cyclical budget deficit the part of the budget deficit which rises in the downswing of the economic cycle and falls in the upswing of the cycle.

structural budget deficit the part of the budget deficit which is not affected by the economic cycle but results from structural change in the economy affecting the government's finances, and also from long-term government policy decisions.

To understand fully the links between the government's budgetary position and the wider economy, it is useful to distinguish between the cyclical and the structural components of the budget deficit and borrowing requirement. The **cyclical budget deficit** is related to the level of aggregate demand in the different phases of the economic cycle. In a boom, when the economy is above its potential, tax receipts are relatively high and spending on unemployment benefit is low. The cyclical deficit falls in the boom, possibly moving into surplus. The reverse happens in the downswing in the economic cycle as tax revenues fall but public spending on unemployment and poverty-related welfare benefits increases. Government borrowing increases and the cyclical deficit gets larger.

The **structural budget deficit** is not related to the state of the economy. This part of the deficit does not disappear when the economy recovers. It thus gives a better guide to the underlying level of the deficit than the headline or cyclical figure. The structural deficit cannot be directly measured, so it has to be estimated.

As the name suggests, growth in the structural component of the budget deficit and borrowing requirement relates in part to the changing structure of the UK economy. (It also relates to government policy decisions: for example, those related to defence expenditure.) In recent years, a number of factors and trends have contributed to the growth of the structural budget deficit. These range from deindustrialisation and globalisation eroding the tax base, via the movement of industries to eastern Europe and Asia, through to an ageing population and the growth of single-parent families dependent on welfare benefits.

A growing structural deficit carries the rather depressing message that a government that seriously wishes to improve public-sector finances will need to introduce significant tax increases or public spending cuts, or possibly both. This is what happened at the time of the rapid deterioration in the UK's public finances from 2008 onwards.

SYNOPTIC LINK

Chapter 15 looks at the problems confronting developing economies. Think about how a country that does not have a comprehensive system of welfare or tax collection is prone to experiencing volatile fluctuations in economic activity.

The consequences of budget deficits and surpluses for macroeconomic performance

The Keynesian view

From the 1950s to 1979, in the UK and in some other countries, mainly in western Europe and also the USA, Keynesian-inspired governments used discretionary macroeconomic policy, in the manner we described earlier, to manage the level of aggregate demand in the economy.

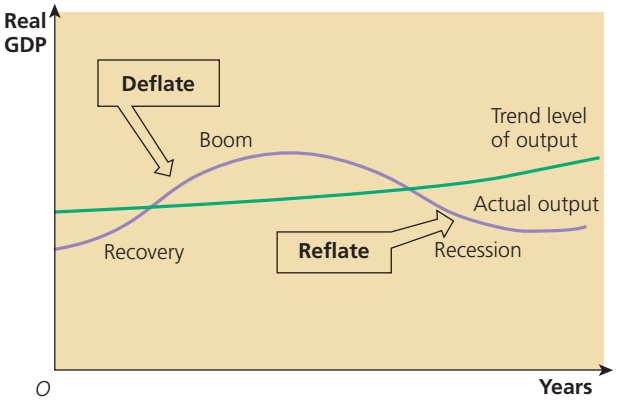


Figure 13.10 Deflating and reflating to 'smooth' the economic cycle

They did this by using contractionary fiscal and/or monetary policy to deflate aggregate demand in the boom phase of the economic cycle (see Figure 13.10), following this with expansionary fiscal and/or monetary policy to reflate (increase) aggregate demand to counter the downswing of the cycle. If successful, the economic cycle would be smoother and less volatile than would be the case had discretionary demand management not been used. And through better utilisation of labour and other resources throughout the cycle, the long-run trend rate of growth might improve. This is known as counter-cyclical demand management policy.

Automatic stabilisers

Reading through our coverage of fiscal policy, you might have concluded that the fiscal policy choice facing a government lies between Keynesian-style discretionary demand management and the 'balancing the budget' approach advocated by many pro-free-market supply-side economists. But in reality, there is an alternative approach that lies between these extremes, in which a government bases fiscal policy on the operation of **automatic stabilisers**. These dampen or reduce the multiplier effects resulting from any change in aggregate demand within the economy, thereby reducing the volatility of the 'ups' and 'downs' in the economic cycle. These 'ups' and 'downs' are illustrated in Figure 13.11. The upper panel of the diagram shows economic cycles in an economy in which there are no automatic stabilisers. The lower panel depicts less volatile cycles brought about by the operation of automatic stabilisers.

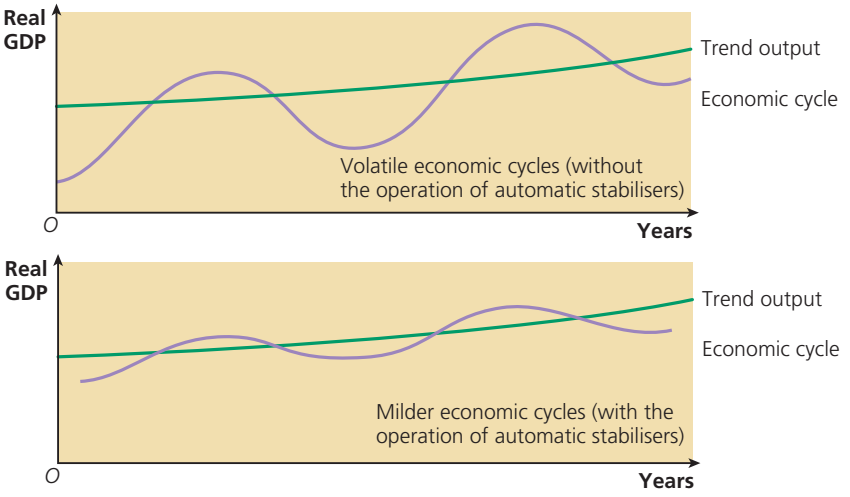


Figure 13.11 The shape of the economic cycle before and after the operation of effective automatic stabilisers

KEY TERM

automatic stabilisers fiscal policy instruments (e.g. progressive taxes and income-related welfare benefits) that automatically stimulate aggregate demand in an economic downswing and depress aggregate demand in an upswing, thereby 'smoothing' the economic cycle.

Suppose, for example, that a collapse of confidence or export orders causes aggregate demand to fall. National income then also begins to fall, declining by the initial fall in demand. But as national income falls and unemployment rises, *demand-led* public spending on unemployment pay and welfare benefits also rises. If the income tax system is progressive, the government's tax revenues fall faster than national income. In this way, increased public spending on transfers and declining tax revenues inject demand back into the economy, thereby stabilising and dampening the deflationary impact of the initial fall in aggregate demand, and reducing the overall size of the contractionary multiplier effect.

Automatic stabilisers also operate in the opposite direction to dampen the expansionary effect of an increase in aggregate demand. As incomes and employment rise, the take-up of means-tested welfare benefits and unemployment-related benefits automatically falls, while at the same time tax revenues rise faster than income. By taking demand out of the economy and reducing the size of the expansionary multiplier, automatic stabilisers reduce overheating in the boom phase of the economic cycle.

TEST YOURSELF 13.9
How might a period of strong economic growth affect the government's fiscal position?

Developed economies and automatic stabilisers

It is now widely agreed that automatic stabilisers such as progressive taxation and income-related transfers contributed to milder economic cycles experienced by the UK, prior to 2008 at least. Before 1939, economic cycles — or trade cycles, as they were then known — were much more volatile, displaying greater fluctuations between boom and slump than in the years between the Second World War and 1973. Keynesians claimed that the relatively mild economic cycles prior to 1973 are evidence of the success of Keynesian demand management policies in stabilising cyclical fluctuations.

However, the economic cycle was relatively mild both in the UK and in countries such as West Germany, which did not use fiscal policy to manage aggregate demand in a discretionary way. This could suggest that the automatic stabilisers of progressive taxation and the safety net provided by welfare benefits for the poor — both of which were introduced widely in western industrialised economies after 1945 — were more significant than discretionary fiscal policy in reducing fluctuations in the economic cycle.

Either way, most economists now agree that a deficit should grow in the downswing of the economic cycle, provided the deficit is matched by a surplus in the subsequent upswing.

There is some doubt, however, as to whether the chancellor at the time — George Osborne — accepted this view when starting and then continuing severe cuts in public spending following the crisis of 2007–08. Arguably, the Osborne cuts, continued by his successor Philip Hammond, created what is now called 'austerity Britain', one feature of which has been major cuts to social care and services.

Although the prime era of 'austerity' appeared to come to an end after the 2015 election, some believed that a new set of spending cuts would be imposed as a way of clearing the deficit caused largely by the Covid-19 pandemic support. However, any reductions in government spending in the period after 2021 are expected to be mild in comparison to those a decade earlier.

The free-market view

Pro-free-market or anti-Keynesian economists are against the use of Keynesian counter-cyclical demand management policies. They argue, in the first place, that with discretionary demand management, governments may get their timing wrong — deflating when they should be reflating, and vice versa. This outcome might be especially likely if governments reflate (i.e. expand) aggregate demand to win votes in a general election, when economic conditions suggest that they should resist this temptation.

The time lags between the change in the state of the economy and when the policy becomes effective also cause problems for policy-makers wishing to fine-tune the economy. For example, if a downturn occurs, the government may implement tax cuts and increases in government spending, but the economy may have recovered on its own by the time these policies become effective.

Free-market economists believe that our knowledge of the economy is imperfect and that in the medium term market forces will drive the economy towards its normal capacity level of output, without government intervention. They believe that government intervention is likely to contribute to greater cyclical instability rather than reduce it: in other words, intervention increases the variability in economic growth over the economic cycle rather than reducing it.

In the second place, whether or not governments get their timing wrong, pro-free-market economists argue that when discretionary fiscal policy is used to try to 'smooth' the economic cycle, the long-run trend rate of growth is likely to fall.

Thirdly, and perhaps most importantly, free-market economists believe that the Keynesian use of demand management policies is just an excuse for the growth of 'big government'. This leads to 'crowding out', in which 'wealth-consuming' government spending replaces the profit motive and 'wealth-creating' private-sector activity. (At the extreme, this is a rather silly argument, as it implies that casinos are 'good' because they are in the private sector, but NHS hospitals are 'bad' because they are in the public sector.)

Financial crowding out

As explained earlier, resource crowding out occurs when resources used by the government cannot simultaneously be employed in the private sector. However, in the context of the possible harmful effects of large budget deficits, *financial* crowding out is more relevant.

Financial crowding out results from the method of financing an increase in public spending. Public spending can be financed by taxation or borrowing. Taxation obviously reduces the spending power of the private individuals and firms paying the taxes. Suppose, however, that the government increases public spending by £40 billion, financing the resulting budget deficit with a sale of new bonds on the capital market, in the form of new sales of gilt-edged securities (gilts). In order to persuade insurance companies, pension funds and the other financial institutions in the market for gilts to buy the extra debt, the guaranteed annual interest rate offered on new gilt issues must increase. But the resulting general rise in interest rates makes it more expensive for firms to borrow and to raise capital. Private-sector investment thus falls and financial crowding out has taken place. Higher interest rates are also likely to reduce household consumption.

TEST YOURSELF 13.10

With the help of *AD/AS* diagrams, explain the difference between supply-side and Keynesian views of the effect of an increase in aggregate demand on the economy.

TEST YOURSELF 13.11
Explain two reasons why a government will run a budget deficit to enhance the productive capacity of an economy.

Benefits and costs of budget deficits

When analysing the benefits and costs of budget deficits, Megan McArdle, a free-market economist, believes that we must again distinguish between the effects of cyclical and structural deficits. According to McArdle, Keynesian economists focus on cyclical deficits, which are not a big problem, whereas pro-free-market economists are more concerned about structural deficits, which can be a big problem.

When correcting cyclical fluctuation in the economy, government intervention may be effective, whether the intervention is based on discretionary demand management or on the automatic fiscal stabilisers we described above. McArdle argues that this intervention can be thought of as 'Great Depression insurance'. In addition, the resulting debt is likely to be eroded by inflation, so it will cost even less in real than in nominal terms. Cyclical deficits should also disappear as the economy recovers; in a boom, a cyclical surplus should emerge, allowing some of the debt to be paid off.

The problem is the structural deficit: the mismatch between government spending and tax revenues that is unaffected by the phase of the economic cycle. The mismatch is manageable as long as the resulting growth of the national debt is roughly the same as, or less than, the rate of GDP growth. In this situation, even persistent structural deficits can be tolerated because the debt to GDP ratio will not increase. But when the structural deficit begins to exceed the rate at which the economy is growing, the economy rapidly runs into trouble. Interest payments begin to grow as a proportion of the overall budget, and as they get bigger, the size of the tax increases or spending cuts needed to close the budget deficit starts to grow. Higher taxes may reduce incentives, and cuts in spending will often harm the most vulnerable.

So what about budget surpluses? On the plus side, budget surpluses may reduce inflationary pressures by taking demand out of the economy, and, as we have explained, they enable the national debt to fall. However, persistent budget surpluses may induce harmful deflation caused by excessive depression of aggregate demand. They can also mean that taxes are higher than necessary, reducing potential growth by damaging the supply-side performance of the economy.

TEST YOURSELF 13.12
Explain why D provides the correct answer to the following question, and why A, B and C do not.

In which one of the following situations is a government most likely to pursue a contractionary fiscal policy in order to decrease aggregate demand?

A When there is a negative output gap

B When the long-run trend rate of economic growth is too high

C When there is a low level of structural unemployment

D When inflationary pressures are operating in the economy

TEST YOURSELF 13.13
Distinguish between cyclical and structural budget deficits.

The significance of the size of the national debt



Government borrowing to finance infrastructure projects like the HS2 rail line can be considered reproductive debt

Some economists and politicians argue that a large national debt is a burden on the economy because future generations of taxpayers end up paying interest on extravagant borrowing by current governments. To assess the merits of this argument, it is useful to distinguish between the *reproductive* national debt and the *deadweight* national debt. We will now look at these terms in more detail.

Suppose that the government sells gilts in order to finance the building of a motorway or some other capital investment or infrastructure project. Although the government is borrowing for a period of many years, the resulting liability is matched by a wealth-producing asset, the motorway. This type of 'reproductive' borrowing is not such a burden on future generations, since interest payments on the debt are in essence 'paid for' out of the motorway's contribution to future national output. (However, some capital assets such as schools and hospitals do not last 'for ever', whereas the debt incurred when they were built often does.)

In contrast, any long-term borrowing to finance *current* spending (e.g. on wars, the salaries of public-sector employees or cash welfare benefits) can be regarded as a burden on future generations, whose taxes will be required to pay interest on the 'deadweight' spending indulged in by the government today. Since the deadweight debt does not cover any real asset, interest payments end up being a burden on both current and future generations of taxpayers. Historically, however, the UK national debt grew fastest during the two world wars of the twentieth century. On the one hand, spending on arms led to a massive growth in deadweight debt. But on the other hand, because the wars were won, the growth of the national debt saved the country from national enslavement. Despite the huge interest payments the UK would have to pay for many decades, arguably this was a price worth paying.

The significance of the national debt is in part dependent on the size of debt as a percentage of GDP, and on whether the national debt can be 'rolled over'. We will look at these in turn.

Debt as a percentage of GDP

If the rate of inflation is greater than the rate at which the budget deficit and borrowing requirement add to the nominal national debt, the money value of the debt as a proportion of money or nominal GDP falls. And if the rate of inflation is greater than the nominal interest rate which the government pays to debt-holders, the government gains and debt-holders lose. This is an example of inflation redistributing wealth from lenders (holders of the national debt) to borrowers (the government), thereby reducing the real burden of the debt to the government. But if holders of the national debt catch on to this, they will try to demand much higher nominal interest payments as a condition for continuing to lend to the government.

STUDY TIP

Don't confuse the terms 'deficit' and 'debt'. Make sure that you can calculate a budget deficit as a proportion of the national debt, and both a budget deficit and the national debt as a proportion of GDP.

SYNOPTIC LINK
Governments finance budget deficits by borrowing. Chapter 12 explained how governments borrow money in the financial markets by selling bonds and gilts.

Rolling over the national debt

Each year, part of the national debt matures and, unless there is a budget surplus, the government must also sell new debt in order to raise the funds with which to repay the maturing debt. This is called renewing or 'rolling over' the national debt. If the government is forced to pay higher nominal interest rates to provide lenders with a real return on their savings, this can lead to the financial 'crowding out' that we described earlier.

The cost of *servicing* the national debt — represented by the interest payments to debt-holders — would not be a problem, first if the national debt were small, and second if all the debt were held *internally* by people living in the country. When the debt is internally held, servicing costs, apart from management costs, are really just a transfer from taxpayers (whose taxes are higher than they would be in the absence of debt interest payments) to debt-holders who have lent to the government. But when the debt is externally held by people living in other countries, the servicing burden of the national debt becomes significant for the country.

The Office for Budget Responsibility

The Office for Budget Responsibility (OBR) was created in 2010 to provide independent analysis of the UK's public-sector finances. The OBR produces medium-term forecasts of the UK economy twice a year in its *Economic and Fiscal Outlook*. By contrast, the Treasury compiles a monthly list of external economic forecasts, comparing them to those of the OBR. Instead of the chancellor making judgements based on Treasury forecasts, the OBR rules on whether the government's policy has a better than 50% chance of meeting the Treasury's fiscal targets.

APPLICATION OF ECONOMICS IN THE REAL WORLD 13.2

The OBR as a fiscal watchdog

In 2012, Robert Chote, the OBR's chairman, explained the OBR's role as a 'fiscal watchdog':

> The OBR is one of a new wave of independent fiscal watchdogs created in recent years. Academics have for some time thought that such bodies could help take some of the politics out of tax and spending policy in the same way that independent central banks had taken much of the politics out of monetary policy.
>
> The core analytical argument for fiscal watchdogs is that, left to their own devices, democratic governments are prone to 'deficit bias' and 'pro-cyclicality' in their management of the public finances.
>
> There are many possible reasons for such bias. Ministers may be seduced by their own rhetoric. Governments may be less forward-looking than voters, driven by elections and impatience. Finance ministries may be weaker than the large spending departments they are meant to control. You might expect outside scrutiny by unofficial bodies (such as the IFS) to be sufficient to counteract these tendencies. But governments can all too easily dismiss the scepticism of outside bodies by pointing out that government ministers have access to privileged information on the behaviour of tax revenues and public spending that outsiders lack.
>
> The creation of the OBR was designed to make the government more accountable for its implementation of fiscal policy. We have been given a statutory entitlement to all the relevant information available within government necessary to fulfil our core duty — 'to examine and report on the sustainability of the public finances'.

In a high-profile speech in 2013, the then prime minister David Cameron said that, according to the OBR, there is 'no alternative' to the government's fiscal austerity programme. But the next day, the OBR published a letter sent to the prime minister taking exception to his claims. A spokesperson at

the OBR said that this shows that the OBR is truly independent and a fiscal watchdog rather than a fiscal lapdog.

Follow-up questions

1 Define the terms 'deficit bias' and 'pro-cyclicality'.
2 Explain why it is desirable to take politics out of tax and public spending policy.
3 The OBR has claimed that it is a 'fiscal watchdog' rather than a 'fiscal lapdog'. What do these terms mean?
4 To what extent, if any, does an independent fiscal watchdog improve fiscal policy? Justify your reasoning.

SECTION 13.1 SUMMARY

- Fiscal policy and supply-side policy provide a way of managing the national economy.
- Fiscal policy uses government spending, taxation and the budgetary position to try to achieve the government's economic policy objectives.
- Keynesian fiscal policy (or demand-side fiscal policy) manages the level of aggregate demand.
- Changes in the government's budget deficit or surplus are important in Keynesian fiscal policy.
- Budget deficits and surpluses are *flow* concepts, and are the difference between the *flows* of government spending and tax revenue.
- The national debt is a *stock* concept, and is the historical accumulation of central government borrowing which has not yet been paid back or redeemed.
- The size of the government spending multiplier affects the power of Keynesian fiscal policy.
- Government spending is a component of aggregate demand.
- Changes in government spending and/or taxation shift the *AD* curve.
- The effect on real output and employment then depends on the shape and slope of the *SRAS* curve — that is, the amount of spare capacity in the economy.

13.2 Supply-side policies

Supply-side economics

Supply-side policies can best be understood if you first understand the meaning of **supply-side economics**. Back in 1983, Arthur Laffer, an eminent supply-side economist, wrote:

> Supply-side economics provides a framework of analysis which relies on personal and private incentives. When incentives change, people's behaviour changes in response. People are attracted towards positive incentives and repelled by the negative. The role of government in such a framework is carried out by the ability of government to alter incentives and thereby affect society's behaviour.

KEY TERM
supply-side policies government economic policies which aim to make markets more competitive and efficient, increase production potential, and shift the *LRAS* curve to the right. Supply-side fiscal policy is arguably the most important type of supply-side policy, but there are also non-fiscal supply-side policies.

KEY TERM
supply-side economics a branch of free-market economics arguing that government policy should be used to improve the competitiveness and efficiency of markets and, through this, the performance of the economy.

Supply-side economics and the free-market revival

Supply-side economics grew in significance in the 1980s as a part of the free-market revival. Free-market economists believe in the virtues of capitalism and competitive markets — a belief which is matched by a distrust and dislike of 'big government' and state intervention in the economy.

The original meaning of supply-side economic policy

When supply-side economics first came to prominence around 1980, it focused narrowly on the effects of fiscal policy on the economy. (Soon after, other supply-side policies were started, including privatisation and trade union reform.) As we have seen, during the Keynesian era most economists regarded fiscal policy — and especially taxation — as a demand management tool. In Keynesian economics, the government's budget deficit lay at the centre of fiscal policy. The Keynesians largely ignored the impact of public spending and tax changes on the supply side of the economy, focusing instead on how changes in government spending and taxation affect aggregate demand.

STUDY TIP
It is important to understand that, originally, supply-side economics was part of the free-market revival. It was meant to provide an alternative way of reducing unemployment and increasing economic growth to the way(s) suggested by Keynesian demand management.

By contrast, supply-side economics initially grew out of the concern expressed by free-market economists about the microeconomic effects of demand-side Keynesian fiscal policy. Indeed, in many respects, supply-side economics is a revival of the old pro-free-market theory that largely disappeared from view during the Keynesian era. The central idea of supply-side economics is that a tax cut should be used, not to stimulate aggregate demand Keynesian-style, but to create incentives by altering relative prices, particularly those of labour and leisure, in favour of work, saving and investment, and entrepreneurship, and against the voluntary choice of unemployment.

The wider meaning of supply-side economic policy

Supply-side economic policy now encompasses more than just fiscal policy; it is the set of government policies which aim to change the underlying structure of the economy and improve the economic performance of markets and industries, and of individual firms and workers within markets. For the most part, supply-side policies are also *microeconomic* rather than simply *macroeconomic*, since, by acting on the motivation and efficiency of individual consumers, workers and entrepreneurs within the economy, the policies aim to enhance general economic performance and the economy's underlying production potential by improving microeconomic incentives.

KEY TERM
supply-side improvements reforms undertaken by the private sector to increase productivity so as to reduce costs and to become more efficient and competitive. Supply-side improvement often results from more investment and innovation, often undertaken by firms without prompting from the government.

STUDY TIP
You should appreciate that many supply-side policies are microeconomic rather than macroeconomic.

Supply-side improvements

Economics students often confuse **supply-side improvement** (or supply-side reform) with supply-side policies. While the two are linked, they do not mean exactly the same thing.

- Supply-side improvements are undertaken by the private sector itself as a result of entrepreneurs realising they must make their firms more efficient

and competitive, first to survive in the modern global economy, and secondly to make profits.

- The government's supply-side policies are part of the means of achieving this desirable outcome.
- In the free-market view, by financially propping up uncompetitive firms, interventionist supply-side policies are often counterproductive in that they do not bring about supply-side improvements.
- This is not necessarily true, however, in the case of interventionist policies which provide training, infrastructure and other external economies that reduce firms' costs.
- Pro-free-market economists generally prefer non-interventionist supply-side policies. By liberalising markets and setting them free, the government pursues the role of *enabler* rather than *provider*.

The difference between free-market and interventionist supply-side policies

Supply-side economists, and free-market economists in general, believe that if markets are allowed to function competitively, the economy is usually close to full employment. However, due to distortions and inefficiencies resulting from Keynesian neglect of the supply side, towards the end of the Keynesian era economic growth and full employment were not achieved. To increase levels of output and employment (and to reduce unemployment), supply-side economists recommend the use of appropriate *microeconomic* policies to remove distortions, improve incentives and generally make markets more competitive.

During the Keynesian era, government microeconomic policy in the UK was generally interventionist, extending the roles of the state and of the planning mechanism. **Interventionist supply-side policies**, such as regional policy, competition policy and industrial relations policy (which were known collectively as industrial policy), generally increased the role of the state and limited the role of markets. The main interventionist supply-side policies supported by free-market economists are government provision of external economies that benefit private-sector firms. These policies include government provision of education and training, and investment in infrastructure projects such as motorways and high-speed trains.

By contrast, pro-free-market supply-side microeconomic policy is *anti*-interventionist, attempting to roll back government interference in the activities of markets and of private economic agents, and to change the economic function of government from *provider* to *enabler*. **Market-based supply-side policies** (or **non-interventionist supply-side policies**) include tax cuts to create incentives to work, save and invest, cuts in welfare benefits to reduce the incentive to choose unemployment rather than a low-paid work alternative, **privatisation**, **marketisation** (**commercialisation**) and **deregulation**.

In essence, the supply-siders, together with the other free-market economists, want to create an enterprise culture. In this broad interpretation, supply-side policies aim to promote entrepreneurship and popular capitalism and to replace the dependency culture that supply-side economists argue had been

KEY TERMS

interventionist policies occur when the government intervenes in, and sometimes replaces, free markets. Interventionist supply-side policies include government funding of research and development.

market-based supply-side policies (or **non-interventionist supply-side policies**) these policies free up markets, promote competition and greater efficiency, and reduce the economic role of the state.

privatisation involves shifting ownership of state-owned assets to the private sector.

marketisation (also known as **commercialisation**) involves shifting provision of goods or services from the non-market sector to the market sector.

deregulation involves removing previously imposed regulations. It is the opposite of regulation.

part of the Keynesian mixed economy. They believe that by creating incentives for businesses and workers, the potential output of the economy will increase and the underlying rate of economic growth will improve. Successful supply-side policies will in turn reduce both unemployment and inflation in the long term, and improve UK external performance, as reflected in the balance of payments on current account. This is because the supply-side reform produced by liberating markets is expected to produce conditions in which domestically produced goods are both price competitive and quality competitive in overseas markets.

However, to bring about the long-run improvements in economic performance required to achieve these ends, substantial and sustainable increases in labour productivity are required. This in turn requires successful reform of the supply side of the economy.

Earlier in the chapter, in the section 'Using an *AD/AS* diagram to illustrate supply-side fiscal policy' (page 470), we described briefly, with the help of Figure 13.3, how supply-side fiscal policy can be used to try to increase the potential output of an economy and hence its long-run rate of economic growth. At this point you should go back and read this section again.

SYNOPTIC LINK

Interventionist supply-side policies are used to correct many of the market failures described in Chapter 8.

The Laffer curve

Supply-side economists believe that high rates of income tax and the overall tax burden create disincentives, which, by reducing national income as taxation increases, also reduce the government's total tax revenue. This effect is illustrated by a Laffer curve, such as the one in Figure 13.12.

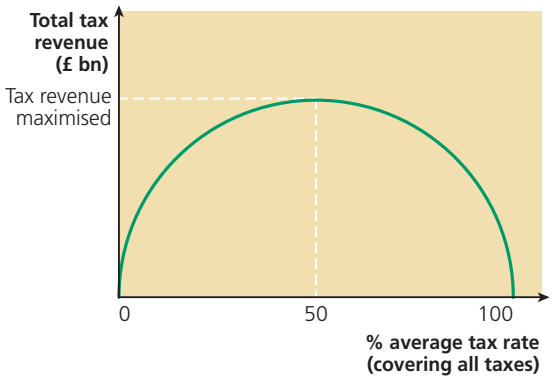


Figure 13.12 A Laffer curve

The Laffer curve, named after the leading supply-side economist Arthur Laffer, quoted earlier, shows how the government's total tax revenue changes as the average tax rate increases from 0% to 100%. Tax revenue must be zero when the average tax rate is 0%, but Figure 13.12 also shows that total tax revenue is assumed to be zero when the tax rate is 100%. With the average tax rate set at 100%, all income must be paid as tax to the government. In this situation,

there is no incentive to produce output other than for subsistence, so with no output produced, the government ends up collecting no tax revenue.

Between the limiting tax rates of 0% and 100%, the Laffer curve shows tax revenue first rising and then falling as the average rate of taxation increases. Tax revenue is maximised at the highest point on the Laffer curve, which in Figure 13.12 occurs at an average tax rate of 50%. Beyond this point, any further increase in the average tax rate becomes counterproductive, causing total tax revenue to fall. (The tax rate which maximises total tax revenue is not necessarily 50%. It can vary through time.)

Supply-side economists argue that the increase in the tax burden in the Keynesian era, needed to finance the growth of the government and the public sector, raised the average tax rate towards or beyond the critical point on the Laffer curve at which tax revenue is maximised. In this situation, any further tax increase has the perverse effect of reducing the government's total tax revenue. Indeed, according to supply-side theory, if the government wishes to increase total tax revenue, it must cut tax rates rather than increase them.

Recent evidence suggests that most governments in developed economies have already benefited from this. As a result, most economies in the Global North (see page 514) would find themselves positioned on the left-hand side of the Laffer curve. Any benefits of further tax cuts may not be significant as a supply-side boost but would simply increase aggregate demand.

A reduction in tax rates creates the incentives needed to stimulate economic growth. Faster growth means that total tax revenue increases despite the fact that tax rates are lower. Arguably, the effect is reinforced by a decline in tax evasion and avoidance, as the incentive to engage in these activities reduces at lower tax rates.

TEST YOURSELF 13.14
Assess the view that a cut in income tax rates inevitably leads to an increase in total tax revenue.

The role of supply-side policies in achieving the government's macroeconomic objectives

Chapter 11 on the measurement of macroeconomic performance introduced you to the government's macroeconomic performance and described how Keynesian and pro-free-market economists have different views on how best to achieve policy objectives such as full employment and control of inflation. At the risk of over-simplifying, Keynesian economists used fiscal policy, and sometimes monetary policy, to manage the level of aggregate demand. At times when it became impossible to achieve all the desired objectives simultaneously — and when policy conflicts arose — the Keynesians 'traded off' between objectives, but they still relied for the most part on demand-side economic policies.

However, Keynesian intervention went into decline when there was a simultaneous failure to achieve any of the prime policy objectives through demand management. Keynesian policies thus gave way to the free-market revival, which eventually evolved into supply-side economics.

Along with many other free-market economists, many if not most supply-side economists believe that government intervention in the economy should be minimised. The aim of policy should be to set markets free and to promote entrepreneurship, in the belief that competitive markets, low taxes, the profit motive and self-interest will then create economic growth. This creates the conditions in which other objectives such as full employment are 'naturally' achieved.

The role of supply-side policies in reducing the natural rate of unemployment

The three panels of Figure 13.13 bring together three of the main elements of supply-side macroeconomics. The upper panel of the diagram depicts the economy's aggregate labour market, while the middle and lower panels respectively show the long-run vertical Phillips curve and long-run *AS* curve.

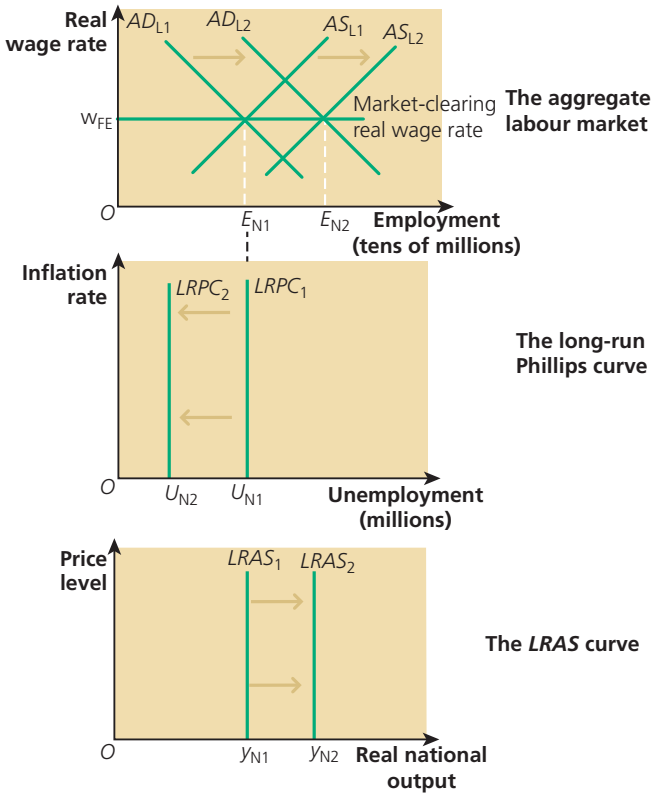


Figure 13.13 Bringing together the natural levels of employment and unemployment and *AD/AS* analysis in supply-side analysis

The upper and lower panels of Figure 13.13 respectively show equilibrium in the economy's aggregate labour market and in the economy's aggregate goods or product market. As the upper panel shows, the natural or 'equilibrium' level of employment E_{N1} is determined in the economy's aggregate labour market at real wage rate w_{FE}. When the labour market is in equilibrium, the number of workers willing to work equals the number of workers whom firms wish to hire.

The middle panel of the diagram shows the natural *level* of unemployment U_{N1} (or natural *rate* if stated as a percentage of the labour force) immediately below E_{N1} in the upper panel. Likewise, the natural level of real output (y_{N1}), which is the same as the 'normal capacity' level of real output, is shown in the lower panel of the diagram, where the long-run aggregate supply curve ($LRAS_1$) is positioned immediately below U_{N1} (in the middle panel) and E_{N1} (in the top panel).

Having set up the three panels of the diagram in Figure 13.13, we can now analyse how, according to the supply-side view, tax cuts can affect employment, unemployment and the 'normal capacity' level of real output in the economy. Suppose the government cuts *employers'* national insurance contributions. Employment costs fall and it becomes more attractive for firms to employ labour. As a result, in the upper panel of Figure 13.13, the aggregate demand for labour curve shifts *rightwards* from AD_{L1} to AD_{L2}.

Likewise, if income tax is cut, this will increase workers' disposable income and thus the incentive to work, shifting the aggregate supply curve of labour *rightwards* from AS_{L1} to AS_{L2}. Both these changes increase the economy's natural level of *employment*, which rises from E_{N1} to E_{N2}.

In the middle panel of Figure 13.13, the long-run Phillips curve shifts *leftwards* from $LRPC_1$ to $LRPC_2$, which means that the economy's natural level of *unemployment* falls from U_{N1} to U_{N2}. Finally, the *increase* in the natural level of employment and the *fall* in the natural level of unemployment, depicted in the top two panels of the diagram, lead to a *rightward* shift of the long-run aggregate supply (*LRAS*) curve in the lower panel of the diagram. The curve shifts from $LRAS_1$ to $LRAS_2$, which means the economy can produce more output without running into the wall of inflation.

Monetary policy and supply-side policies

When supply-side policies were first implemented about 40 years ago, extreme supply-side economists argued that *neither* monetary policy *nor* fiscal policy should be used to manage aggregate demand. Instead, government activity should be restricted to creating the supply-side conditions (and 'sound' money) in which markets function competitively and efficiently.

Most free-market economists now accept that monetary policy should be used to manage demand, but that it should be supportive, both of the supply-side reforms introduced over the years, and of any further supply-side changes to be made in the future. The essential tasks of monetary policy are to make sure there is just enough demand to absorb the extra output produced by the supply side of the economy, and to maintain the general public's confidence in 'sound' money by achieving a low, stable inflation rate.

UK governments (and the Bank of England) now realise that, because monetary and fiscal policy are interdependent rather than independent of each other, the success of monetary policy depends on the fiscal policy implemented by the Treasury. As we have seen, whenever it runs a budget deficit, the government has to borrow to finance the difference between spending and revenue. Conversely, a budget surplus enables repayment of past borrowing and a fall in the national debt. As we explained in Chapter 12, governments borrow in two main ways, by selling short-dated debt or Treasury bills, and by borrowing long term by selling government bonds or gilts. Treasury bills are sold by the government to the banks, which create new bank deposits

when they buy the bills. As bank deposits are a form of money, this type of government borrowing increases the money supply and makes it difficult for monetary policy to control monetary growth.

By contrast, new issues of gilts are largely sold to non-bank financial institutions such as pension funds and insurance companies. To persuade these institutions to finance a growing budget deficit, the government may have to raise the rate of interest offered on new gilt issues. But higher interest rates discourage investment in capital goods by private-sector firms. (This is the financial crowding-out process referred to earlier in this chapter.)

Either way, government spending financed by borrowing has implications for monetary policy and produces an arguably undesirable result. Either the borrowing increases the money supply (which may be inflationary), or it raises interest rates (which may crowd out private-sector investment). The undesirable monetary consequences resulting from budget deficits and government borrowing explain why UK governments now try to make fiscal policy consistent with, and subordinate to, the needs of monetary policy.

SYNOPTIC LINK
Go back to Chapter 11 and reread the explanations of the natural rate of unemployment (NRU) and the 'normal capacity' level of real output or GDP, on pages 391 and 397–98. This shows why policies to increase *AD* will not always reduce unemployment.

Supply-side policies, the rate of inflation and the balance of payments

Supporters of the use of pro-free-market supply-side policies argue that, by reducing businesses' costs of production and by reducing monopoly profits through making markets more price competitive, supply-side policies reduce cost-push inflationary pressures. By enabling productive capacity to grow in line with aggregate demand, successful supply-side policies also help to reduce demand-pull inflation pressures.

By creating 'lean and fit' firms, successful supply-side policies and supply-side reforms are also likely to improve the country's quality competitiveness. Highly competitive firms, which invest in product design and 'state-of-the-art' technology and methods of production, end up producing high-quality goods which customers demand. Taken together, increased price and quality competitiveness give the country a competitive edge, in both domestic and overseas markets. With domestic customers switching to high-quality home-produced goods and overseas residents buying more of the country's exports, the country's balance of payments on current account should improve.

TEST YOURSELF 13.15
Explain how successful supply-side policies can improve the UK's balance of payments position on the current account.

SYNOPTIC LINK
By improving the economic performance of the domestic economy, supply-side policies aim to increase the international competitiveness of the country's exports, and thereby improve the balance of payments on current account (see Chapter 14).

EXTENSION MATERIAL

Expansionary fiscal contractionism

Figure 13.14 illustrates how the then chancellor George Osborne implemented fiscal policy in the period immediately after May 2010. Osborne believed that severe cuts in government spending would free resources for the private sector to use, and that the private sector would make better use of these resources than the government. If this was the case, workers who lost their jobs in the public sector would soon be employed more productively in the private sector. This view of how fiscal policy should operate has been called 'expansionary fiscal contractionism' (EFC) or 'expansionary austerity'. The expansionary fiscal contractionism hypothesis predicts that, given certain assumptions about how people behave, a major reduction in government spending that changes future expectations about taxes and government spending can expand private-sector spending, resulting in overall economic expansion.

Not all economists accept the EFC hypothesis. Christina Romer at the University of California commented:

> Despite what I feel is overwhelming and compelling evidence that fiscal stimulus is expansionary, and fiscal contraction is, well, contractionary, many politicians claim the opposite is true ... But even more striking are the number who assert forcefully that fiscal austerity — getting the budget deficit down immediately — would be good for unemployment and growth.

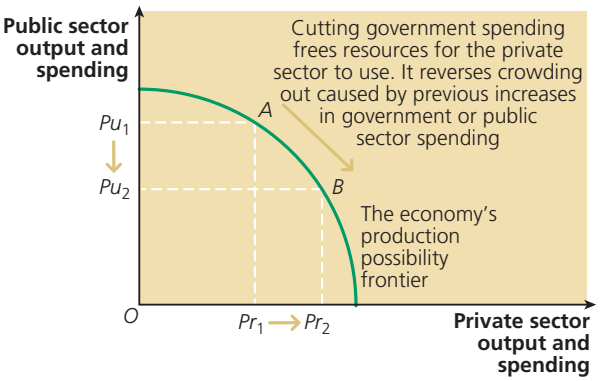


Figure 13.14 Expansionary fiscal contractionism

The 'trickle-down' effect

Along with other free-market economists, supply-side economists believe that, in the long run, expansionary fiscal policy leads to inflation, with no increase in real output. Some believe that although the rich benefit most from supply-side tax cuts, a 'trickle-down' effect means that the working poor also benefit. This is because the rich respond to tax cuts by employing more members of staff, albeit generally on low wages.

But other economists question the strength and even the existence of trickle-down effects. There is little evidence to support the existence of trickle-down. There is also an increased likelihood that the rich will move their money out of the country into a country where taxes on incomes and wealth are lower — which would mean the trickle-down effect is less likely to occur. And even if a trickle-down effect does operate, the widening income inequalities that are responsible for more of the poor being employed by the rich are viewed by Keynesian economists as far too inequitable or unfair.

APPLICATION OF ECONOMICS IN THE REAL WORLD 13.3

Have UK politicians been influenced by supply-side theory?

On 18 December 2012, the then chancellor George Osborne spoke to a gathering of American supply-side economists at the Manhattan Institute for Policy Research in New York City. In answer to a question about the division between American Republicans and British Conservatives on the power of tax rate cuts to generate tax revenue, George Osborne replied:

> Well I'm a fiscal Conservative, and I don't want to take risks with my public finances on an assumption that we are at some point on the Laffer curve. What I would say is, let's see the proof in the pudding: in

other words, I'm a low-tax Conservative, I want to reduce taxes, but I basically think you have to do the hard work of reducing government spending to pay for those lower taxes. If you want to cut taxes, cut welfare and cut spending, and that's what I'm doing.

Osborne seemed to be rejecting the 'extreme' supply-side argument, based on the Laffer curve, that tax cuts can be self-financing and that they are all that is needed for growth generation. Osborne, along with other Conservative members of the UK government between 2010 and 2016, was best seen as a 'moderate' rather than an 'extreme' supply-sider.

Many centrist Labour and Liberal Democrat politicians now also accept moderate supply-side arguments, though they continue to reject extreme supply-side calls for severe tax cuts and greater income inequality in order to incentivise the population. Especially since the emergence of problems caused by a large national debt, few politicians now call for continuous large budget deficits as the way to achieve growth and full employment, and there is much agreement that the tax structure should be used in a supply-side way to create incentives for work, entrepreneurship, saving and investment.

Follow-up questions

1 Explain, using the Laffer curve, how a government may aim to increase tax revenues by cutting taxes.
2 Describe the difference between 'extreme' and 'moderate' supply-side views on the role of tax cuts.
3 Which groups in society will be negatively affected by large cuts in public spending? Explain your answer.
4 Explain how the growth of the national debt has affected fiscal policy.

EXTENSION MATERIAL

Microeconomic supply-side policies

So far, for the most part, we have been describing and explaining the macroeconomic effect of supply-side policies. This can give the misleading impression that supply-side economics and policies are mainly macroeconomic. However, the opposite is closer to the truth: most supply-side policies are based on microeconomic fundamentals such as the roles of self-interest and entrepreneurship.

The supply-side theory of the effects of taxation on labour market incentives, which lies at the heart of free-market supply-side economics, depends on the shape of the supply curve of labour.

Supply-side economists usually assume a conventional upward-sloping supply curve of labour. Such a curve, which is illustrated in Figure 13.15(a) on the next page, shows that workers respond to higher wage rates by supplying more labour.

Since a cut in the rate at which income tax rates are levied is equivalent to an increase in the wage rate, the upward-sloping supply curve implies that workers respond to cuts in the marginal rate of income tax by working harder. (The marginal rate of income tax is the percentage of the last pound of income paid in tax.) If this is the case, a reduction in income tax rates creates the incentive for workers to supply more labour (and for entrepreneurs to become more enterprising), while an increase in income tax rates has a disincentive effect on effort and the supply of labour.

However, the supply curve of labour need not necessarily slope upwards throughout its length. The backward-bending labour supply curve in Figure 13.15(b) is another possibility. It shows that, above the hourly wage rate W_1, any further wage rate increase (or income tax decrease) causes workers to supply *less* rather than *more* labour. In this situation, workers prefer to enjoy extra leisure time rather than to work. Following an increase in the hourly wage rate from W_1 to W_2, the hours of labour time supplied fall from L_1 to L_2.

It is important to note that, if supply curves of labour bend backwards, the supply-side argument — that tax reductions increase national output and efficiency through their effect on labour market incentives — becomes much weaker. Far from encouraging people to work harder, a wage rise or income tax cut might have the opposite effect, causing people to work fewer hours and to enjoy more leisure time instead.

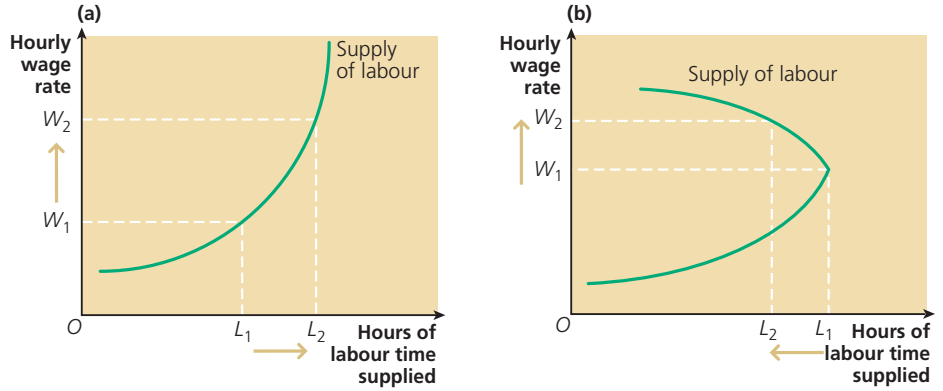


Figure 13.15 Microeconomic labour supply curves

STUDY TIP

The backward-bending labour supply curve is not included on the current AQA specification but can still be referred to in terms of both labour markets and supply-side policies if appropriate.

Examples of microeconomic supply-side policies

Industrial policy measures

- **Privatisation.** This involves the sale or transfer of assets such as nationalised industries from the public sector to the private sector.
- **Marketisation (or commercialisation).** This means shifting economic activity from non-market provision (financed by taxation) to commercial or market provision for which the customer pays.
- **Deregulation.** This involves the removal of previously imposed regulations in order to promote competition. Deregulation removes barriers to market entry to make markets contestable (i.e. there is a greater potential for competition), and gets rid of unnecessary bureaucracy, which had increased firms' costs of production.
- **Internal markets.** In the National Health Service and education, where the state continues to be a major producer and provider of services, internal markets can be introduced to provide a form of commercial discipline and to improve efficiency. In an internal market, which is a substitute for privatisation, the taxpayer continues to finance hospitals and schools, but hospitals and schools 'earn' the money according to how many patients and pupils they attract.
- **Subsidising spending on research and development.** This often takes the form of a cut in corporation tax rates to increase business profits. This provides firms with greater financial resources to boost investment and increase spending on research and development, which should lead to better products and higher sales in the long run.

SYNOPTIC LINK

Chapter 8 explains how privatisation, marketisation and deregulation seek to increase levels of competition in markets and make firms more efficient.

Labour market measures

- **Lower rates of income tax.** Marginal rates of income tax can be reduced to create labour market incentives, and tax thresholds or personal tax allowances raised to remove the low-paid from the tax net.
- **Reducing state welfare benefits relative to average earnings.** Lower benefit levels create incentives to choose low-paid employment in preference to claiming unemployment-related benefits. In addition, welfare benefits can be made more difficult to claim, and available only to claimants genuinely looking for work. Making benefits less attractive may also reduce the unemployment trap.
- **Changing employment law to reduce the power of trade unions.** Possible changes include: removing trade unions' legal protection, restricting their rights, and extending the freedom for workers not to belong to unions, and for employers not to recognise and negotiate with unions; replacing collective bargaining with individual wage negotiation and employer determination of pay; and restricting the right to strike and to undertake industrial action.
- **Repealing legislation which limits employers' freedom to employ.** This makes it easier for employers to 'hire and fire' workers.
- **More flexible pension arrangements.** Workers may be encouraged to 'opt out' of state pensions and to arrange private pension plans so as to reduce the burden on taxpayers. They might be allowed to transfer private-sector pensions between employers when changing jobs.
- **Improving the training of labour.** This might involve establishing training agencies and academies to develop vocational technical education. However, UK governments have rejected the proposal to impose a 'training tax' on all employers to prevent free-riding by firms with no training schemes, which poach trained workers from firms that do train their workers.
- **Short-term employment contracts.** Though not a government policy, the introduction of short-term employment contracts in labour markets replaced 'jobs for life' with short-term labour contracts. This, together with the promotion of profit-related and performance-related pay, has had a supply-side effect. Zero-hours employment contracts (where a worker has a contract that guarantees no hours of work — but the worker may be used) have been introduced by a large number of employers. In 2013 just under 600,000 workers in the UK were employed on zero-hours contracts. This has risen to around 1 million since 2020 (although more than two-thirds of those on zero-hours contracts are part-time workers). Critics of these policies believe they lead to even greater poverty and inequality for ordinary workers in an increasingly casualised and exploited part-time labour force.

SYNOPTIC LINK

You should refer back to Chapter 6 on the labour market to review the theories of how labour markets operate.

Financial and capital market measures

- **Deregulating financial markets.** Financial deregulation has created greater competition among banks and building societies, and opened up the UK financial markets to overseas banks and financial institutions. These reforms have increased the supply of funds and reduced the cost of borrowing for

UK firms. Financial deregulation and the removal of foreign exchange controls have also encouraged 'inward' investment by overseas firms such as Samsung and Nissan. However, very lightly regulated banks created by financial deregulation contributed significantly to the financial crisis which began in 2007 and then ushered in the 2008–09 recession. Regulation since then has been tightened — though some would argue, not sufficiently.

- **Encouraging saving.** Governments have created special tax privileges for saving. They also encouraged saving by giving individual shareholders first preference in the market for shares issued when former nationalised industries such as British Gas were privatised. However, most individual shareholders quickly sold their shares to institutional shareholders, which negated one of the main reasons for privatisation. It is also worth noting that, in recent years, very low interest rates, brought about by the government's monetary policy, have discouraged saving. To help savers, the amount of interest that can be earned before any tax is paid on it has been lifted significantly and now nearly all savers pay no tax on their savings' interest, although higher interest rates have reduced the number receiving tax-free interest.
- **Promoting entrepreneurship.** Governments have encouraged the growth of popular capitalism and an enterprise culture. Company taxation has been reduced and markets have been deregulated to encourage risk taking.
- **Reducing public spending and public-sector borrowing.** This is intended to free resources for private-sector use and avoid crowding out.

A further look at market-based versus interventionist supply-side policies

The supply-side policies that we have described so far in this chapter have generally been market-based, in the sense that they have been used to 'free up' markets, to 'roll back' the economic functions of the state, and to replace public-sector economic activity with that of the private sector. Free-market economists seek to promote market conditions by designing policies which stimulate market forces in three main ways:

- tax cuts to create incentives to seek profits and work harder
- privatisation to unleash market forces by taking businesses out of the public sector
- deregulation to encourage enterprise and increase competition.

This view of the supply side of the economy stems from the original use of the term 'supply-side economics' in the free-market revival 40 years ago.

However, interventionist policies can also target the supply side, provided that they aim to improve the efficiency and economic performance of individual firms, industries and markets. Interventionist supply-side policies include:

- government spending on education and training
- industrial and regional policy
- spending on infrastructure such as motorways
- subsidising spending on research and development (R&D).

Interventionist supply-side policies are based on the view that government intervention in the economy is needed to correct market failure and short-termism, which may reduce investment and growth. More recent pro-free-market supply-side policies, in contrast, have generally aimed to correct and reduce government failure.

SECTION 13.2 SUMMARY

- Supply-side fiscal policy affects the position of the *LRAS* curve.
- In supply-side fiscal policy, tax changes are used to try to change incentives in the economy.
- Free-market supply-side economists argue that government policy should be used to improve incentives and the competitiveness and efficiency of markets.
- In its early years, supply-side economics focused on how, via increased incentives, tax cuts promote economic growth and are self-financing.
- The growth of supply-side economics has been part of the free-market revival.
- Supply-side economists are generally anti-interventionist and wish to reduce the economic role of the state. However, some supply-side policies, such as government financing of retraining schemes in labour markets, are interventionist.
- Supply-side policies contribute to supply-side reform and supply-side improvements.
- Most economists now accept the argument that the supply side of the economy is just as important as the demand side.
- Many supply-side policies are microeconomic rather than macroeconomic.

Questions

1 Explain the relationship between a budget deficit and the national debt.

2 Explain the role of fiscal policy in supply-side economic policy.

3 To what extent do you agree that supply-side improvements are needed if UK macroeconomic performance is to improve? Justify your answer.

4 Explain the difference between the cyclical and the structural components of a budget deficit.

5 Explain how automatic stabilisers operate.

6 Evaluate the impact of measures undertaken by governments to reduce the budget deficit.

7 Do you agree that supply-side policies on their own are sufficient to help the economy achieve a stable, sustainable and satisfactory rate of economic growth? Justify your answer.



14 The international economy: globalisation and international trade

Although we have made brief reference to aspects of the international economy in earlier chapters, particularly those on macroeconomics, we have not explained international economics in any systematic way. The purpose of the two remaining chapters in the book is to explain the key elements of the international economy and to examine their impact on the UK economy. This chapter looks first at globalisation, and then at trade and the balance of payments. This is followed in Chapter 15 by an explanation of exchange rate systems and a discussion of various international issues related to economic growth and the development of low-income countries.

LEARNING OBJECTIVES

These are to understand:

- the meaning of globalisation and its main elements
- the basis of international trade and the principle of comparative advantage
- the arguments for protectionism and the different forms of protectionist policies
- patterns of trade, both globally and for the UK
- the main features of a customs union, the Single European Market (SEM)
- the role of the World Trade Organization (WTO)
- the different sections of the balance of payments account and the relationship between them
- the significance of balance of payments deficits and surpluses
- how deflation, import controls, devaluation and supply-side measures and policies can be used to try to correct a current account imbalance.

14.1 Globalisation

The meaning of globalisation

Globalisation is the name given to the processes that integrate all or most of the world's economies, making countries increasingly dependent upon each other. Some economists argue that globalisation has occurred over centuries, going back at least as far as the creation of a system of relatively free worldwide trade in the nineteenth century. Perhaps it extends even further back to some European countries' occupation of other parts of the world.

KEY TERM
globalisation the process of increasing economic integration of the world's economies.

In the late nineteenth century and the period before 1914, communication and transport networks expanded throughout much of the world and international trade grew significantly. At the same time, older industrial countries, and particularly the UK, began to invest their surplus savings in capital projects located overseas rather than in their domestic economies.

However, these changes are better described as aspects of *internationalisation* rather than *globalisation*. Globalisation, which has come to mean rather more than mere internationalisation of economic relationships, began to feature in the economics literature of the mid-1980s. The use of the term has increased dramatically ever since.

The causes of globalisation

Recent globalisation has been made possible by improvements in information and communication technology (ICT), as well as by developments in transport and other more traditional forms of technology. Examples of globalisation include service industries in the UK dealing with customers through call centres in India, and fashion companies designing their products in Europe, making them in southeast Asia and finally selling most of them in North America.

Indeed, this textbook illustrates globalisation. The text on page ii of this book states that it is typeset in India and printed in Bosnia & Herzegovina, while being written and published in England.

The main characteristics of globalisation

Some of the main characteristics of globalisation are:

- the growth of international trade and the reduction of trade barriers, known as trade liberalisation — a process encouraged by the **World Trade Organization** (WTO)
- greater international mobility of both capital and labour
- a significant increase in the power of international capitalism and **multinational corporations** (MNCs) or transnational companies
- the deindustrialisation of older industrial regions and countries, and the movement of manufacturing industries to newly industrialising countries (NICs)
- more recently, the movement of internationally mobile service industries, such as call centres and accounts offices, to NICs
- a decrease in governmental power to influence decisions made by MNCs to shift economic activity between countries.

KEY TERMS
World Trade Organization an international body whose purpose is to promote free trade by persuading countries to abolish import tariffs and other barriers to trade. As such, it has become closely associated with globalisation.

multinational corporations enterprises operating in several countries but with their headquarters in one country.

TEST YOURSELF 14.1
Explain how reducing trade barriers has made product markets more contestable (see Chapter 5, section 5.10 for more details).

Figure 14.1 depicts some of these characteristics or features of globalisation in the form of a 'mind map'.

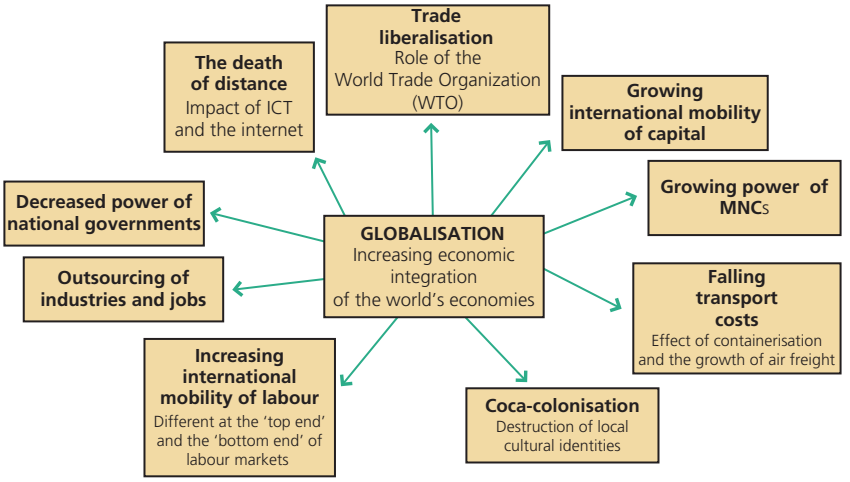


Figure 14.1 Some of the main characteristics or features of globalisation

The consequences of globalisation for less developed economies

KEY TERM
less developed economies countries considered behind in terms of their economy, human capital, infrastructure and industrial base.

Less developed economies are characterised by low national income per head, a high rate of population growth, low levels of human capital, high unemployment, poor infrastructure and over-dependence on the export of a few primary commodities. Many nations in Asia, Africa and Latin America fit this model.

For the critics of globalisation, low-paid workers in 'sweatshops' (factories or production facilities where workers are subject to poor working conditions — long hours, limited checks on workers' welfare — and very low wages), farmers in the developing world being forced to grow genetically modified (GM) crops, the privatisation of state-owned industry in order to qualify for International Monetary Fund (IMF) and World Bank loans, and the growing dominance of US corporate culture and multinational companies symbolise what is wrong with globalisation in its effects on less developed economies.



Multinationals often use lower-paid workers in developing countries to manufacture their products

According to this view, globalisation has led to a 'McDonaldisation' or 'Coca-colonisation' of significant parts of the world's economy, which involves the destruction of local and national products, identities and cultures by US world brands. This process is to some extent offset by a process known as 'glocalisation'. This word, which is a combination of the words 'globalisation' and 'localisation', reflects the idea that a global product or service is more likely to succeed if it is adapted to local practices and cultural expectations. An example of glocalisation is McDonald's restaurants offering menus to satisfy local tastes in different countries. For example, in parts of the world where beef is not eaten widely, the 'Maharaja Mac' — a burger made of chicken — is promoted as the equivalent of the 'Big Mac' in beef-eating countries. In this sense, glocalisation is an extension of globalisation, being another way of extending the reach of global brands.

STUDY TIP

You should understand how different groups in an economy will be affected in different ways by an event or policy. Identify how globalisation has affected different groups in both developed and emerging economies over the last 30 years. Make a list of the costs and benefits of globalisation and use examples to support your line of reasoning.

Closely related to the 'world brand' process, there has been controversy concerning the treatment of local labour by multinational corporations. On the one hand, companies such as Nike are accused of selling trainers and footballs in developed economies such as the UK, at prices far above the cost of raw materials and the low wages paid to workers in developing economies who make the goods. On the other hand, the multinationals argue that the 'low wages' they pay far exceed the local wages paid by indigenous firms. They believe this encourages local wages to rise. MNCs also claim to improve labour productivity, health and safety, and other labour market conditions in the poor countries in which they operate.

The consequences of globalisation for more developed economies

The anti-globalisation argument that 'Coca-colonisation' of significant parts of the world's economy has had a harmful effect on the economies of less developed economies is not accepted by supporters of globalisation, who believe that people in the rest of the world demand the products supplied by global corporations because they consider them superior to traditional local produce.

KEY TERM

more developed economies countries with a high degree of economic development, high average income per head, high standards of living, usually service industries dominating manufacturing, and investment having taken place over many years in human capital and infrastructure.

The anti-globalisation lobby also believe that, by threatening to close down factories and offices in **more developed economies** and to move production to poor countries, MNCs may reduce wages and living standards in developed economies. The significance of this depends on the type of jobs that emerge in developed economies to replace those lost through deindustrialisation and globalisation. The new jobs being created may be in the highly skilled service sector, but may also be menial, low-paid, unskilled jobs.

The arguments used so far in this and the previous section have been cast somewhat in a 'lose/win' or 'win/lose' format, implying that winners gain at the expense of losers. However, supporters of globalisation who believe that the world as a whole is a net winner from the globalisation process argue that globalisation stems in part from the massive increase in world trade that has taken place since the Second World War. This itself has only been possible because of increased trade liberalisation and the reduction of import controls and other forms of protectionism. They believe that free trade leads to a better allocation of resources and that additional benefits have been derived from international competition and the spread of technology making most people better off.

Dependency theory of trade and development

The belief that globalisation benefits most countries — less developed economies as well as more developed economies — has been attacked by supporters of the *dependency theory* of trade and development. Dependency

theorists argue that many developing economies possess little capital because the system of world trade and payments has been organised by developed industrial economies to their own advantage. The terms of trade — the ratio of a country's export prices to its import prices — have generally moved in favour of industrialised countries and against primary producers. This means that, by exporting the same amount of manufactured goods to developing economies, a developed economy can import a greater quantity of raw materials or foodstuffs in exchange.

By the same token, the developing economy must export more in order to buy the same quantity of capital goods or energy vital for development. Globally, the movement of the terms of trade in favour of developed economies has raised levels of income and standards of living in the richer economies at the expense of poorer developing economies. However, there have been exceptions, such as the oil-producing non-industrial countries, which have benefited from substantial increases in the price of oil over recent decades.

Economists of the dependency school argue that the transfer of wealth and resources to the richer economies is further promoted by profit flows and interest payments. On an international scale, dividends and profits flow to multinational corporations with headquarters in North America, western Europe and Japan from their subsidiaries in the developing world. Similarly, there is a flow of interest payments to western banks from loans originally made to finance development in developing economies. In most years, flows of dividends and interest payments from the Global South to the Global North exceed aid flows in the opposite direction. (The countries of the Global North include the USA, the UK, Germany, France and Japan, and all other countries benefiting from a high degree of economic development. The Global North gets its name from the fact that most of the countries which comprise the Global North are geographically located to the north of most of the countries of the Global South. The latter are poorer developing economies located in parts of the world such as sub-Saharan Africa.)

STUDY TIP

Use appropriate diagrams to explain how investment by an MNC can affect the economy of a less economically developed economy. For example, a diagram showing a production possibility frontier shifting outwards can be used to explain how the economy may experience economic growth, whereas a negative production externality can illustrate how pollution may be dumped on third parties in society. Remember always to integrate diagrams into written answers.

SYNOPTIC LINK

See page 533 for an explanation of 'Global North–Global North' and 'Global North–Global South' patterns of world trade.

Is globalisation good or bad?

Free-market economists generally support globalisation and regard its growth as inevitable. They argue that the benefits of further global economic integration, which include the extension of political freedom and democracy as

well as the economic benefits of more production and higher living standards, significantly exceed the disadvantages, such as the destruction of local cultures. However, opponents argue that globalisation is a respectable name for the growing exploitation of the poor, mostly in developing economies, by international capitalism and US economic and cultural imperialism.

Globalisation in the service sector

KEY TERM
European Union an economic and partially political union established in 1993. It developed as a more closely integrated organisation from the European Economic Community (EEC), originally formed in 1957. The EU has expanded to include (after the UK's departure) 27 countries.

Until quite recently, it was said that manufacturing was much more internationally mobile than service-sector employment, but this is now disputed. Call centres became one of the fastest-growing sources of employment in the UK in the 1980s and 1990s. At that time, UK-based companies favoured locating call centres in regions of high unemployment (and relatively low wages) within the UK. To some extent this has now changed. Many call centres and back-office activities of firms in industries such as financial services have been moved to less developed economies, notably on the Indian subcontinent, and also to the eastern member countries of the **European Union** (EU), such as the Czech Republic. These movements result from the 'death of distance' (which means that distance becomes less of a barrier for businesses and consumers), which is an important part of the globalisation process. The rapid development of electronic methods of communication means that many service activities can now be located anywhere in the world, with little or no effect on a company's ability to provide the service efficiently to its customers.

Four factors encouraging the overseas location of call centres have been:

- relatively low wages in developing economies and in eastern and central Europe
- highly reliable and cheap telecommunications
- 24-hour shift employment to overcome the problem of time zones
- workers fluent in English, which is now the world's business language.

However, for call centres, a fifth factor is often lacking: many overseas workers are insufficiently familiar with UK culture and habits, which leads to a communication problem. This has recently led to some call centres being relocated back to the UK. This disadvantage is much less significant for back-office employment: for example, employing people in India to administer a UK company's accounts.

Global labour and capital mobility

As the previous paragraphs indicate, globalisation involves moving capital to lower-cost labour much more than it involves allowing low-paid workers born in poor countries to enter rich countries in North America and Europe. However, since the late nineteenth century there has been a much greater movement of poor people into rich countries than ever before. To some extent, immigration controls introduced by countries such as the USA and Australia, which replaced an earlier completely free movement of labour, have slowed this process. But this has been offset by illegal immigration and the fact that rich countries informally encourage migration to fill the relatively low-paid jobs that their own citizens do not wish to do. MNCs recruit skilled labour from other countries and governments also encourage highly trained and talented individuals to migrate to fill skill shortages.

A larger EU (especially after the 2004 expansion, when a further 10 countries joined) is increasing both labour and capital mobility on a regional basis. Western European firms have been moving eastwards, but this is balanced by workers from countries such as Poland and Hungary migrating westwards. Nonetheless, it is still much easier in a globalised world for a brain surgeon or a highly paid business executive to move between countries than it is for a low-skilled worker.

Globalisation and the power of national governments

In recent decades, globalisation has considerably reduced the power of national governments, certainly in smaller countries, to control multinational firms operating within their boundaries. National governments have also lost much of the freedom to undertake the economic policies of their choice with respect to managing domestic economies. Governments enjoy less freedom to introduce tariffs and other import controls. At the same time, capital flows into and out of currencies severely constrain a government's ability to implement an independent monetary policy, even when the country's exchange rate is freely floating.

SYNOPTIC LINK

Chapter 15 looks at exchange rate systems and explains how central banks can set monetary policy to manage the value of their country's currency. It also explains why many countries have a floating exchange rate.

The role of multinational corporations in globalisation

As explained earlier, multinational corporations (MNCs) are business enterprises operating in several countries to manage production or to deliver goods and services, but with their headquarters in just one country. The growth of MNCs since the 1970s has driven globalisation in four main ways.

Economic integration and increased trade

MNCs have built global production platforms which enable specialisation and take advantage of the global division of labour. They have broken the production process down so that specialist workers in different geographical locations contribute value to the development of products. For example, the inscription on the back of Apple products states: 'Designed by Apple in California. Assembled in China.' The high-skilled development takes place in Silicon Valley in California, but the low-skilled assembly is undertaken where labour is significantly cheaper.

By organising global production platforms and supply chains which stretch across international boundaries, MNCs have increased the economic integration of different countries. Products are no longer made in one country and then exported across the world. Instead components are made in a variety of locations and assembled in different regional factories. The technological revolution in communication and transportation has allowed MNCs to coordinate global business activities and open up new markets. In turn, trade

flows — the buying and selling of imports and exports between countries — have increased as MNCs have moved products along their global supply chains.

Investment and technology transfers

MNCs think globally and seek to locate operations in countries that offer the best business opportunities. In the late 1990s the oil company British Petroleum changed its name to BP to reflect a new global identity. MNCs invest and create jobs in regions that appear attractive. By creating a business-friendly environment, since the 1990s the Chinese government has attracted foreign direct investment (covered later in this chapter) worth billions of dollars from MNCs. The Chinese currency has remained relatively weak against the US dollar and the abundant supply of cheap labour has meant that firms which have located low- and medium-skilled manufacturing in China have been able to reduce production costs significantly. Since 1990, hundreds of millions of Chinese workers have been lifted out of poverty as they have moved from subsistence farming to better-paid factory jobs.

The long-term investment flows undertaken by MNCs have seen technology transferred from developed economies such as the USA and the UK to the emerging market economies, most notably China and India. Capital equipment and advanced technology that was once located in western economies has enabled countries in southeast Asia to industrialise. In the long term, economic power is shifting from West to East. However, this has caused tensions, most notably between the US and Chinese governments. American firms have been unhappy with a lack of legal protection, and in some cases theft of intellectual property and the forced transfers of technology to China. US businesses have been forced to hand over proprietary technology in exchange for access to the Chinese mainland markets. This has eroded the technological advantages of US firms and allowed Chinese businesses to develop highly technical products without incurring research and development costs.

Forced transfers of technology were an important part of the trade war that started in 2018 between the USA and China. A trade war is where countries take steps to restrict the flow of imports from one country or region, often provoking retaliation from the other country. Former President Donald Trump argued that forced technology transfers amounted to illegal state aid by the Chinese government. Even with the election of Joe Biden as president in 2020, the USA–China trade war continues.

Changing employment patterns and global capitalism

The process of globalisation has seen MNCs transform employment patterns across the world. The long-term trend has been for manufacturing and industrial production to close down in the advanced western economies and move to the emerging market economies in southeast Asia. This has resulted in structural unemployment and large pockets of poverty in the industrial regions of North America and Europe, but the creation of hundreds of millions of jobs, lifting workers out of absolute poverty, in China and India. MNCs have coordinated this process.

SYNOPTIC LINK

Structural unemployment was covered in Chapter 11. Although it is connected to geographic and occupational immobility, the loss of jobs in domestic industries due to low-cost foreign competition in low-wage economies is a further explanation for structural unemployment.

MNCs are profit-maximising enterprises and have used their leverage to demand a favourable business environment when deciding where to locate factories and offices. For decades, governments across the world have been under pressure to cut corporate taxes, reduce social protection and workers' rights legislation, resist environmental protection and create a light-touch system of regulation. Globalisation has been driven by the demands of MNCs, which have affected inequality in two different ways. First, many of the poorest workers in the world have benefited from higher incomes as they have taken up jobs in factories. The higher incomes have reduced the level of inequality between the developed world and the emerging markets. This has, however, had its costs, such as environmental degradation, pollution, long hours and dangerous workplaces. Second, a significant number of the workers in the developed world have experienced stagnating living standards and lower real incomes as a result of well-paid employment disappearing and relocating to cheaper labour zones. In these economies, inequality has increased.

The global marketplace and international brands

MNCs compete against each other in markets around the world. They seek to operate a global scale of production which enables them to benefit from economies of scale and to reduce average production costs significantly. MNCs develop brands which are recognised across the world, although products are often tailored to suit the tastes of particular markets. Standardised products have eroded national boundaries. Consumer choice has widened, in the sense that a greater variety of global products are now available, but it has decreased as a result of local firms either disappearing or being assimilated into multinational conglomerates owned elsewhere. Globalisation means that MNCs sell to consumers in different continents the same clothes, cars, computers, phones, food and beverages.

TEST YOURSELF 14.2

MNCs invest in less economically developed countries (LEDCs) and can increase employment, capital equipment, access to technology, pollution and environmental destruction, and worker exploitation.

To what extent does investment by an MNC in an LEDC increase living standards?

APPLICATION OF ECONOMICS IN THE REAL WORLD 14.1

Economic nationalism and global production platforms

The global car industry is dominated by a few large firms. Modern cars are highly sophisticated products that contain thousands of component parts, from windscreens to satellite navigation systems, to airbags. To build vehicles, car makers have established complex supply chains across international borders so that they can source high-quality parts from specialist suppliers.

The US car makers, General Motors and Ford, have established production platforms across North America to build cars for the lucrative American markets. The now-defunct North American Free Trade Agreement (NAFTA) set out rules that firms selling in the USA, Mexico and Canada had to comply with several conditions to qualify for tariff-free trade. Under NAFTA rules, 62.5% of a vehicle had to be made in the USA, Canada or Mexico. Since the NAFTA trade bloc was established in 1994, US car makers have located their factories in Mexico to take advantage of the lower labour costs. In Europe, German firms have invested in central and eastern Europe, as have Japanese companies in China. This allows firms which invest in this way to bring costs down so as to remain price competitive.

Global production platforms have enabled multinational companies to take advantage of specialisation and the division of labour (see Chapter 4 for more on this) and then to bring prices down for consumers. However, this has enraged economic nationalists such as former US president Donald Trump. He argued that, through MNCs moving factories to Mexico, American communities in the US car industry's traditional heartlands in Michigan have lost thousands of jobs. This has created structural unemployment and large pockets of poverty. Economic nationalists believe that the US government needs to protect American jobs by forcing multinational firms to relocate factories back to the USA.

The Trump administration argued that the NAFTA trade deal disadvantaged US workers. His advisors wanted to renegotiate the trade rules and compel car makers to produce 85% of vehicles in North America and at least 50% in the USA. Trump dislikes multilateral negotiations, favouring instead bilateral treaties because he believes the USA can secure more favourable terms by dealing with countries individually rather than in round-table talks. In discussions with the Mexican government, President Trump demanded the introduction of $16 hourly wage in all car factories across North and Central America to prevent Mexican firms from undercutting American competitors.

The nationalist 'America First' agenda may be popular in the US 'rust belt', which has faced many years of decline, but it is questionable whether the strategy will be successful in the long run. Forcing multinational firms to relocate factories in the USA will raise unit costs in a fiercely competitive industry. Import tariffs can be increased to protect US firms from competition, but consumers will experience higher prices. Furthermore, the managers at the car makers will deploy considerable resources into reorganising supply chains to comply with the new rules. This will divert funds away from projects that seek to improve production methods and develop new technologies. Finally, the $16 an hour wage may appeal to workers but in the long run it is likely to result in higher unemployment. Car firms originally located low-skilled manufacturing in Mexico to take advantage of lower labour costs. But by raising the wages to $16 an hour, firms will be incentivised to replace low-skilled workers with automated machinery.

NAFTA was replaced in 2020 by the United States-Mexico-Canada Agreement (USMCA) and did take into account some of the criticisms made of the original NAFTA deal. Differences between NAFTA and USMCA included more environmental and working regulations. Greater incentives were also included to encourage car production in the USA, and more access was given to the dairy market in Canada.

Follow-up questions

1 Define the term 'supply chain'.
2 Explain two ways in which a government can seek to protect domestic firms from competition.
3 Explain how building a global production platform enables a multinational corporation to reduce costs and increase profits.

APPLICATION OF ECONOMICS IN THE REAL WORLD 14.2

Has globalisation given way to de-globalisation?

Robert J. Samuelson of Investors.com recently asked: 'What has happened to globalisation?' Samuelson argued that, for decades, growing volumes of cross-border trade and money flows have fuelled strong economic growth. But something remarkable is happening — the growth rate of trade and money flows is slowing and, in some cases, declining. Samuelson pondered whether these changes herald prolonged economic stagnation and rising nationalism or, optimistically, whether they make the world economy more stable and politically acceptable.

For workers employed or previously employed in US manufacturing industries, some aspects of de-globalisation are very attractive. Globalisation has sucked factory jobs out of North America. But now, the tide may be turning. Apple has announced a $100 million investment to return some Mac computer manufacturing to the USA. Though small, the decision may reflect a new trend.

General Electric's sprawling Appliance Park in Kentucky once symbolised the USA's manufacturing process, with employment peaking at 23,000 in 1973. Since then, jobs have shifted abroad or succumbed to automation. However, now General Electric (GE) is moving production of water heaters, refrigerators and other appliances back to Appliance Park from China and Mexico. Nor is GE alone. Otis is moving some elevator output from Mexico to South Carolina.

China's labour cost advantage has eroded. By 2020, China's manufacturing costs were projected to be US$6.5 billion. Although wages of US production workers average were much higher, other non-wage factors favour the USA. American workers are more productive; automation has cut labour costs, and cheaper natural gas further lowers costs; and finally, higher oil prices have boosted freight rates for imports.

Follow-up questions

1 What is meant by de-globalisation?
2 How may the growth in 'cross-border trade' promote globalisation?
3 Explain how the US government can use the threat of tariffs to encourage firms to relocate factories in the USA.
4 How may the movement of factories from Mexico to the USA increase the costs of production of multinational corporations?

SECTION 14.1 SUMMARY

- Globalisation is the name given to the increasing integration of the world's economies.
- Trade liberalisation, international capital and labour mobility, and the increased power of multinational corporations (MNCs) are important elements of globalisation.
- MNCs have driven many parts of the globalisation process.
- In recent years there has been an increase in support for economic nationalists who seek to reverse globalisation using protectionist policies.
- De-globalisation has seen MNCs close factories in less developed economies and relocate them in developed economies.

14.2 Trade

Imagine a small country such as Iceland in a world without international trade. As a closed economy, Iceland's production possibilities would be limited to the goods and services that its narrow resource base (i.e. limited range of natural resources) can produce. Iceland's average costs of production would probably be high because its small population and the absence of export markets would mean that economies of scale and long production runs could not be achieved. At the same time, the consumption possibilities of Iceland's inhabitants would be restricted to the goods that the country can produce.

Compare this with Iceland's position in a world completely open to international trade. In an open economy, imports of raw materials and energy greatly boost Iceland's production possibilities. In theory at least, Iceland can now produce a much wider range of goods. In practice, however, Iceland produces the relatively few goods and services that it is good at producing, and imports all the rest. By gaining access to the much larger world market, Iceland's industries benefit from economies of scale and long production runs. Likewise, imports of food and other consumer goods present Iceland's inhabitants with a vast array of choice and the possibility of a much higher living standard and level of economic welfare than would be possible in a world without trade.

While the UK is a much larger country than Iceland, it still has a relatively narrow resource base. The arguments outlined above about using trade to widen production and consumption possibilities help to explain the role of

international trade in the UK economy. The UK's economy is dependent on foreign trade. The country's departure from the EU has clearly been disruptive to the UK economy as the agreement reached between the EU and the UK on trade after 'Brexit' has made trade more difficult and less 'free', despite there being no tariffs on trade in goods.

SYNOPTIC LINK

Chapter 4 explains how firms can reduce average costs by increasing their scale of production and benefit from global economies of scale. Earlier in this chapter, we explained how MNCs have built global production platforms that enable them to operate on a global scale and benefit from global economies of scale. International trade enables consumers to buy products from MNCs and enjoy the benefits of lower average prices.

The model of comparative advantage

We now follow the rather 'common sense' justification of the benefits of international specialisation and trade given above with a more rigorous explanation provided by the principle of comparative advantage.

To show how the principle of *comparative advantage* explains some of the benefits of international specialisation and free trade, it is helpful to look first at a related concept: *absolute advantage*.

Absolute advantage

KEY TERM

absolute advantage a country has an absolute advantage if it can produce more of a good than other countries from the same amount of resources.

A country has an **absolute advantage** if it can produce more of a good with a given amount of resources than another country. (Alternatively, we might say that it can produce the same amount of the product with fewer resources.)

To explain absolute advantage, we will assume just two countries in the world economy, Atlantis and Pacifica. Each country has only 2 units of resource. Only two goods can be produced: guns and butter. They may seem an odd choice of products but they have traditionally often been used as examples in economics. In one way, it does not matter which pair of products is chosen — the models used in trade theory would still illustrate the same points. However, the example of guns and butter is useful in other ways.

Guns can be thought of as a military product, provided by the government, and butter as a consumption product for the general public. How much priority the government places on defining itself through military products and how much it produces products for the immediate well-being of the population can tell you a lot about the government's views overall. You can also think of these in terms of a capital good (guns) that is not for immediate use and may only be used as a deterrent in the future, and a consumption good (butter) which will be used up once consumed and does not add to the long-term productive capacity of the economy.

Each unit of resource, or indeed a fraction of each unit (because we will assume that resources or inputs are divisible), can be switched from one industry to another, if so desired, in each country. Table 14.1 shows each country's production possibilities from 1 unit of resource.

Table 14.1 Production with 1 unit of resource

	Guns		Butter
Atlantis	4	or	2 tonnes
Pacifica	1	or	6 tonnes

Quite clearly, Atlantis has an absolute advantage in producing guns. It is four times better in gun production than Pacifica. However, this is not the case for butter production. Pacifica is three times better in butter production and so possesses an absolute advantage in this good.

Suppose that both countries devote half their total resources to each activity (i.e. 1 unit of resource out of the 2 units available for each country). We will call this *production without specialisation*. Without specialisation, Atlantis produces 4 guns and Pacifica produces 1 gun, which means that 5 guns are produced in total. Likewise, without specialisation, total butter production is 8 tonnes. Atlantis produces 2 tonnes and Pacifica produces 6 tonnes.

Table 14.2 shows this outcome.

Table 14.2 Production without specialisation

	Guns		Butter
Atlantis	4	and	2 tonnes
Pacifica	1	and	6 tonnes
Total output without specialisation	5	and	8 tonnes

Now let's see what happens when each country produces only the good in which it has an absolute advantage. Table 14.3 shows that, if Atlantis devotes both its resource units to guns, it produces 8 guns. Likewise, if Pacifica completely specialises, the country produces 12 tonnes of butter with its 2 units of resource.

Table 14.3 Production with complete specialisation

	Guns		Butter
Atlantis	8	and	0 tonnes
Pacifica	0	and	12 tonnes
Total output with complete specialisation	8	and	12 tonnes

The final table in this series, Table 14.4, shows how, when each country enjoys an absolute advantage in a different good, complete specialisation results in more of both goods being produced.

Table 14.4 Output gain from complete specialisation

	Guns		Butter
Atlantis	8	and	0 tonnes
Pacifica	0	and	12 tonnes
Total output without specialisation	5	and	8 tonnes
Total output with complete specialisation	8	and	12 tonnes
Output gains from specialisation	3	and	4 tonnes

In this example, specialisation produces a net output gain of 3 guns and 4 tonnes of butter. Since more has been produced, more can be consumed, and it is possible to make people better off.

However, for *output* gains to translate into gains from *trade*, two further factors have to be taken into account. First, administration and transport costs occur whenever trade takes place. As a result, the net gains from trade are:

(3 guns + 4 tonnes of butter) – transport and administration costs

Clearly, specialisation and trade are not worthwhile if transport and administration costs exceed the output gains resulting from specialisation.

Secondly, assuming that only two countries trade with each other, for output gains to transfer into *welfare* gains for the inhabitants of both countries, the goods being traded must be in demand in each of the importing countries. Given this assumption about demand, we will further assume that each country exports its surplus to the other country once it has satisfied its own inhabitants' demand for the good in which it specialises. (This *double coincidence of wants* is not necessary when more than two countries trade together, as there will be an increased range of goods and services produced.)

But suppose Atlantis's inhabitants are vegans who refuse to eat animal products, while Pacifica's inhabitants are pacifists who hate guns. For Atlantis's inhabitants, butter is a 'bad' rather than a good. Likewise, guns are a 'bad' for Pacifica's residents. (A good yields utility or economic welfare to consumers, but a bad yields disutility or negative welfare.) Atlantis refuses to import butter, and Pacifica refuses to buy guns. Specialisation and trade do not take place. Without suitable demand conditions, the case for specialisation and trade disappears.

SYNOPTIC LINK

The extension material in Chapter 8, page 232, explains the distinction between 'goods' and 'bads' in the context of public goods and public bads.

Comparative advantage

KEY TERM

comparative advantage this is measured in terms of opportunity cost: the country with the least opportunity cost when producing a good possesses a comparative advantage in that good.

Absolute advantage must not be confused with the rather more subtle concept of **comparative advantage**. However, understanding absolute advantage is a stepping-stone to understanding comparative advantage. To explain comparative advantage, we will change the production possibilities of both countries so that Atlantis possesses the absolute advantage for both guns and butter (which means that Pacifica has an absolute disadvantage in both goods). Table 14.5 shows how much 1 unit of resource can produce in each country.

Table 14.5 Production with 1 unit of resource

	Guns		Butter
Atlantis	4	or	2 tonnes
Pacifica	1	or	1 tonne

Although Atlantis is 'better at' — or has an absolute advantage in — producing both guns and butter, the country possesses a comparative advantage in the production of guns but has a comparative disadvantage in the production of

butter. *This is because comparative advantage is measured in terms of opportunity cost, or what a country gives up when it increases the output of a product by 1 unit.*

The country that gives up least of the other commodity when increasing output of a particular commodity by 1 unit possesses a comparative advantage in that good. Ask yourself how many guns Atlantis has to give up in order to increase its output of butter by 1 tonne. The answer is 2 guns. But Pacifica only has to give up 1 gun to produce an extra tonne of butter. Thus Pacifica possesses a comparative advantage in butter production even though it has an absolute disadvantage in both products.

Table 14.6 shows the total output of guns and butter, if each country devotes 1 unit of resource to each industry.

Table 14.6 Production without specialisation

	Guns		Butter
Atlantis	4	and	2 tonnes
Pacifica	1	and	1 tonne
Total output without specialisation	5	and	3 tonnes

When one country possesses an absolute advantage in both products, as in the example above, its comparative advantage always lies in producing the good in which its absolute advantage is greater. Similarly, the country that has an absolute disadvantage in both activities possesses a comparative advantage in the product in which its absolute disadvantage is less. Table 14.5 shows that Atlantis has a comparative advantage in gun production, while Pacifica's comparative advantage lies in butter production.

Complete specialisation, shown in Table 14.7, results in more guns but less butter being produced. Without specialisation, the combined output of both countries is 5 guns and 3 tonnes of butter (Table 14.5). With complete specialisation, this changes to 8 guns and 2 tonnes of butter. While production of guns has increased, production of butter has fallen.

Table 14.7 Output gain from complete specialisation

	Guns		Butter
Atlantis	8	and	2 tonnes
Pacifica	0	and	1 tonne
Total output without specialisation	5	and	3 tonnes
Total output with complete specialisation	8	and	2 tonnes

When one country has an absolute advantage in both goods, complete specialisation in accordance with the principle of comparative advantage does not result in a net output gain. The output of one good rises, but the output of the other good falls.

STUDY TIP

Remember that the country which possesses an absolute advantage in *both* goods enjoys a comparative advantage in the good in which its absolute advantage is greater. Likewise, the other country, which has an absolute disadvantage in both goods, has a comparative advantage in the good in which its absolute disadvantage is less.

Students are sometimes puzzled by this, thinking (wrongly) that because of the loss of output of one of the two goods, specialisation cannot lead to gains from trade. However, *partial* specialisation can produce a net output gain. For example, suppose Pacifica (which suffers an absolute disadvantage in both goods) specialises completely and produces 2 tonnes of butter. By contrast, Atlantis (which has the absolute advantage in both goods) devotes just enough resource (half a unit) to top up world production of butter to 3 tonnes. This means that Atlantis can still produce 6 guns using 1.5 units of resource.

Total production in both countries is therefore 6 guns and 3 tonnes of butter. At least as much butter and more guns are now produced compared with the 'no specialisation' outcome. This example shows that specialisation can produce a net output gain, even though one country is absolutely better at both activities.

TEST YOURSELF 14.3

Table 14.8 shows the production possibilities of two countries with 1 unit of resource.

Table 14.8 Production with 1 unit of resource

	Guns		Butter
Oceania	20	or	10
Eurasia	15	or	5

Which statement is correct? Explain why.

A Oceania has an absolute advantage in the production of both products so trade will not take place.

B Oceania has a comparative advantage in the production of both guns and butter.

C Eurasia has a comparative advantage in the production of guns.

D Eurasia has a comparative advantage in the production of both guns and butter.

STUDY TIP

Make sure you don't confuse the 'terms of trade' with the 'balance of trade'. The balance of trade is a major part of the balance of payments on current account, which is explained later in this chapter. The balance of trade is the difference in value between a country's imports and exports. If the value of exports exceeds the value of imports, there is a balance of trade surplus. Conversely, if the value of exports is less than the value of imports, there is a balance of trade deficit.

EXTENSION MATERIAL

Adam Smith, David Ricardo and absolute and comparative advantage

Although he did not use the precise term, Adam Smith is credited with introducing the concept of absolute advantage in his great eighteenth-century book, *The Wealth of Nations*. Some years later, in the early nineteenth century, another distinguished classical economist, David Ricardo, developed Adam Smith's ideas into the principle of comparative advantage. Neither Smith nor Ricardo was interested solely in abstract economic theory. Instead, like many great economists, they wished to change society for the better by influencing the politicians of their day.

Smith and Ricardo believed in the virtues of a competitive market economy and industrial capitalism. Ricardo, in particular, believed that a single country such as the UK, and indeed the whole world economy, can only reach its full productive potential, maximising output, welfare and living standards, if the market economy is truly international. He argued that each country should specialise in the activities in which it possesses a comparative advantage, and then trade the output that is surplus to its needs in a world free of tariffs and other forms of protectionism.

The assumptions underlying the principle of comparative advantage

When arguing that definite benefits result when countries specialise and trade in accordance with the principle of comparative advantage, we made a number of rather strong but not necessarily realistic assumptions. Indeed, the case for trade — and hence the case *against* import controls and other forms of protectionism — is heavily dependent upon these assumptions. Likewise, some of the arguments in favour of import controls and against free trade, which we will explain shortly, depend on showing that the assumptions necessary for the benefits of specialisation and trade to occur are simply not met in real life.

The assumptions underlying the principles of comparative advantage are as follows:

- Each country's endowment of factors of production, including capital and labour, is fixed and immobile between countries, although factors can be switched between industries within a country. In the course of international trade, finished goods rather than factors of production or inputs are assumed to be mobile between countries.
- There are *constant returns to scale*. In our example, 1 unit of resource is assumed to produce 4 guns or 2 tonnes of butter in Atlantis, whether it is the first unit of resource employed or the millionth unit. But in the real world, increasing or decreasing returns to scale are both possible and indeed likely. In a world of increasing returns to scale, the more a country specialises in an activity in which it initially has an absolute advantage, the more its productive efficiency and advantage increases. Countries that are 'best' to start with become even 'better'. But if decreasing returns to scale occur, specialisation erodes efficiency and destroys a country's initial advantage. A good example occurs in agriculture, where over-specialisation can result in monoculture or the growing of a single cash crop for export. Monoculture often leads to soil erosion, vulnerability to pests and falling future agricultural yields.
- Demand and cost conditions are relatively stable. Over-specialisation can cause a country to become particularly vulnerable to sudden changes in

demand or to changes in the cost and availability of imported raw materials or energy. Changes in costs, and new inventions and technical progress, can quickly eliminate a country's earlier comparative advantage. The greater the uncertainty about the future, the weaker the case for a country specialising in a narrow range of products.

STUDY TIP

There are other justifications of specialisation and free trade beside the argument that global economic welfare increases if countries specialise and trade in accordance with the principle of comparative advantage. We mentioned a justification earlier in our discussion of widening production possibilities and consumption possibilities. A further justification, also mentioned briefly, is the ability to enjoy the benefits of economies of scale that lead to greater efficiency and lower average costs. (If you remember from Chapter 4, economies and diseconomies of scale are closely linked to increasing returns to scale and decreasing returns to scale.)

EXTENSION MATERIAL

Comparative advantage and competitive advantage

Comparative advantage must not be confused with *competitive* advantage. A country, or a firm within a country, enjoys a competitive advantage when it produces better-quality goods at lower costs and better prices than its rivals. Competitive advantage is more similar to absolute advantage than to comparative advantage.

Dynamic factors that promote the growth of firms can create competitive advantage. Successful investment undertaken over many years equips a country with modern, 'state-of-the-art' production capacity, capable of producing high-quality goods that people want to buy. Properly funded and organised research and development (R&D) contributes in a similar way, while the stock of human capital resulting from investment in education and training adds to competitive advantage.

Factors that create competitive advantage can trigger a virtuous spiral of larger profits, higher investment, better products and greater sales, which in turn leads to even higher profits, and so on. Conversely, countries and firms that are not competitive may enter a vicious spiral of decline. Inability to compete causes profits to fall, which in turn reduces investment. The quality of goods declines and sales are lost to more competitive countries or firms. Profits again fall and this may continue if the country cannot become competitive.

Supporters of free trade would argue that this does not matter. If there is free trade between countries, this creates exactly the circumstances needed to foster the growth of competitive advantage. However, as we will see in the next section, in our coverage of strategic trade theory, others argue that, if used *strategically*, protectionism is better at providing selected industries with a competitive advantage against the rest of the world.

The case for import controls and protectionism

Import controls can be divided into quantity controls such as import **quotas**, which put a maximum limit on imports, and **tariffs** or **import duties** (and their opposite, **export subsidies**), which raise the price of imports (or reduce the price of exports).

Supporters of free trade believe that import controls prevent countries from specialising in activities in which they have a comparative advantage and

KEY TERMS

quotas physical limits on the quantities of imported goods allowed into a country.

tariffs (also known as **import duties**) taxes imposed on imports from other countries entering a country.

export subsidies money given to domestic firms by the government to encourage firms to sell their products abroad and to help make their goods cheaper in export markets.

from trading their surpluses. As a result, production takes place inefficiently and economic welfare is reduced. However, as already noted, the case for free trade depends to a large extent upon some of the assumptions underlying the principle of comparative advantage. Relax these assumptions and the case for free trade is weakened. Nevertheless, it should be remembered that the proponents of free trade also emphasise the dynamic benefits that result from opening up an economy to competition from abroad.

Below are some of the contexts in which import controls have been justified.

STUDY TIP

When considering the case for free trade (or the case for import controls) it is best to avoid a 'shopping list' approach. Instead, select, develop, analyse and evaluate three or four arguments or counter-arguments. Also consider the likelihood of retaliation and its implications.

Infant industries

As already explained, many economic activities benefit from increasing returns to scale, which mean that the more a country specialises in a particular industry, the more productively efficient it becomes. This increases its competitive advantage. Developing economies justify the use of import controls to protect infant industries from established rivals in advanced economies. It is argued that protectionism is needed while newly established industries develop and achieve full economies of scale.

Sunset industries

A rather similar case to the infant industry argument is sometimes made in advanced industrial economies such as the UK to protect older industries from the competition of infant industries in developing economies. Some economists advocate the selective use of import controls as a potentially effective supply-side policy instrument to prevent unnecessary deindustrialisation and to allow orderly rather than disruptive structural change in the manufacturing base of the economy. According to this view, import controls are justified, at least on a temporary basis, to minimise the social and economic cost of the painful adjustment process, as the structure of an economy alters in response either to changing demand or to changing technology and comparative and competitive advantage. Labour is not perfectly mobile and import controls may help to reduce the incidence of structural unemployment.

Strategic trade theory

The infant and sunset industry arguments are closely related to strategic trade theory, a relatively new theory that has grown in influence in recent years. Strategic trade theory argues that comparative and competitive advantage are often not 'natural'. Rather, governments try to create competitive advantage by nurturing strategically selected industries or economic sectors. This justifies protecting the industries while competitive advantage is being built up. The skills that are gained will then spill over to help other sectors in the economy. Strategic trade theory also argues that protectionism can prevent exploitation by a foreign-based monopoly. Governments in more developed economies use two types of strategic trade policy to help declining industries such as textiles and shipbuilding:

- trade adjustment assistance and other aid to workers and firms in these industries
- subsidies on exports or taxes on imports and investment or adjustment assistance subsidies to protect the industries from foreign competition.

However, export subsidies, with few exceptions, such as agricultural products, violate World Trade Organization (WTO) rules. By contrast, less developed economies use strategic trade policies to protect and promote the growth of their infant industries.

Agricultural efficiency

As noted earlier, monoculture, which might result from specialisation and trade, erodes efficiency and destroys comparative advantage that existed before specialisation took place. Decreasing returns to scale weaken the case for complete specialisation.

Changes in demand or cost conditions

Over-specialisation may cause a country to become particularly vulnerable to sudden changes in demand or to changes in the cost and availability of imported raw materials or energy. Specialisation can lead to a lack of diversity in a country's economy. In this situation, countries could benefit from diversifying their economies, and import controls might lead to this outcome.

Anti-dumping

When a country produces too much of a good for its own domestic market, the surplus may then be 'dumped' — sold at a price below cost — in overseas markets. Import controls are often justified as a means to prevent this supposedly 'unfair' competition. Article 6 of the General Agreement on Tariffs and Trade allows countries to protect themselves against dumping where there is a material injury to the competing domestic industry.

Self-sufficiency

Politically, it is often argued that protection is necessary for military and strategic reasons to ensure that a country is relatively self-sufficient in vital foodstuffs, energy and raw materials in a time of war. Many are using this sort of argument in 2022 as countries attempt to reduce reliance on Russian gas as part of the economic sanctions imposed as a result of the war in Ukraine.

Employment

Trade unions argue that import controls are necessary to prevent multinational firms shifting capital to low-wage developing economies and exporting their output back to the countries from which the capital was moved. They further argue the case for employing labour, however inefficiently, in protected industries rather than allowing labour to suffer the greatest inefficiency of all: mass unemployment. This is an example of *second-best theory*. It stems from the fact that the 'first best' (free trade in a world of fully employed economies and perfect markets) is unattainable. Therefore, a country can settle legitimately for the second best. Employing resources perhaps inefficiently, protected by tariffs, is better than leaving resources, such as labour, unemployed. This justification for protecting domestic industries returned to prominence after the financial crisis in 2009.

The 2016 election of Donald Trump as president in the USA saw the country move away from its traditional position as the champion of free trade and

towards economic nationalism and protectionism. The Trump administration started trade disputes with China and long-term allies Canada and Mexico as part of a strategy to bring factories and jobs back to the industrial regions that supported the president's 'America First' agenda. Although President Biden has since adopted a more conciliatory tone, there are no obvious signs of de-escalation in this trade war as of 2022 — although the Covid-19 pandemic and Russia's invasion of Ukraine have diverted attention away from trade disputes.

SYNOPTIC LINK

Unemployment is a significant cost of international trade. Many industries in the UK have closed down because British firms have been unable to compete with foreign competitors. Go back over Chapter 11 and look at how structural unemployment has affected the UK since the 1970s. Pay attention to the increase in UK unemployment in the early 1980s, particularly in the industrial cities in the north of England.



Since the Trump presidency, trade policy has become a source of conflict between the USA and China

STUDY TIP

The case for free trade, or trade liberalisation, is the case against import controls and other forms of protectionism. Likewise the case for import controls is the case against free trade. However, remember that it is not unreasonable to argue that trade liberalisation is generally desirable but to accept that there are circumstances when protectionism may be justified. There is a wide spectrum of opinion among economists about the extent to which protectionist measures are economically justifiable.

APPLICATION OF ECONOMICS IN THE REAL WORLD 14.3

Paul Krugman and strategic trade theory

In 2008, Paul Krugman, professor of economics at Princeton University in the USA, was awarded the Nobel prize in economics, in part for his pioneering work in developing strategic trade theory in the 1980s and 1990s. In contrast to the orthodox free-market approach to international trade, firmly grounded in the principle of comparative advantage, strategic trade theory argues that import controls can sometimes be justified. Krugman argued that rich, developed economies benefited from protectionism while they established their national wealth. However, they now put pressure on poor countries to abandon import controls and to allow overseas-based multinational corporations unlimited access to their economies. As we mentioned earlier in the chapter, opponents of globalisation argue that free trade theory is used to justify first-world economic imperialism.

Below is an extract from the beginning of a paper Krugman wrote in 1987:

> If there were an Economist's Creed, it would surely contain the affirmations 'I understand the principle of comparative advantage' and 'I advocate free trade'. For 170 years, the appreciation that international trade benefits a country whether it is 'fair' or not has been one of the touchstones of professionalism in economics.
>
> Yet the case for free trade is currently more in doubt than at any time since the 1817 publication of Ricardo's *Principles of Political Economy*. This is not because of the political pressures for protection, which have triumphed in the past without shaking the intellectual foundation of comparative advantage theory. Rather, it is because of the changes that have recently taken place in the theory of international trade itself.
>
> In the last 10 years the traditional **constant returns**, **perfect competition** models of international trade have been supplemented and to some extent supplanted by a new breed of models that emphasise **increasing returns** and **imperfect competition**. These new models call into doubt the extent to which actual trade can be explained by comparative advantage.
>
> Showing that free trade is better than no trade is not the same thing as showing that free trade is better than sophisticated government intervention. The view that free trade is the best of all possible policies is part of the general case for laissez-faire in a market economy and rests on the proposition that markets are efficient. If increasing returns and imperfect competition are necessary parts of the explanation of international trade, we are living in a **second-best** world where government intervention can in principle improve on market outcomes.
>
> Source: Paul Krugman, 'Is free trade passé?', *Journal of Economic Perspectives*, 1987.

The full article can be found online at: **https://tinyurl.com/mr2xchp9**

Follow-up questions

1. What is comparative advantage?
2. Briefly define each of the terms in bold in the extract.
3. Do you agree that increasing returns to scale and imperfect competition weaken the case for unlimited free trade based on the principle of comparative advantage? Justify your answer.
4. What is first-world economic imperialism?

The causes and consequences of countries adopting protectionist policies

In this and the next two sections of the chapter, we bypass the justifications for tariffs put forward by economists such as Paul Krugman and return to the free-market case for free trade. Analysis of the welfare gains from free trade and the welfare losses caused by import controls centres on the concepts of consumer surplus and producer surplus that were introduced in Chapter 5.

If a country does not enter into international trade, which means its economy is *closed*, domestic demand for a good within a country can only be met by domestic supply (i.e. by firms producing *within* the country). Such a situation is shown in Figure 14.2, where market equilibrium for the good occurs at point

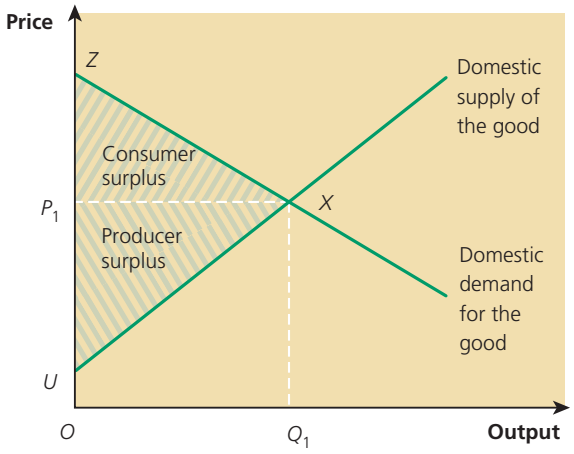


Figure 14.2 Consumer and producer surplus in a market closed to international trade

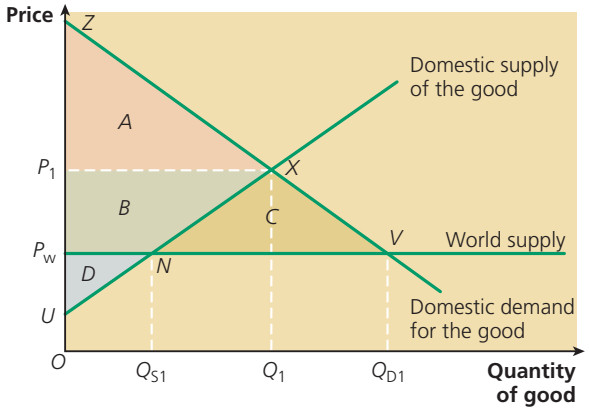


Figure 14.3 The welfare gains and losses resulting in opening up an economy to cheaper foreign imports

X. Consumers pay price P_1 for the good, and the quantity bought and sold is Q_1. Consumer surplus, which is a measure of consumer welfare, is shown by the triangular area bounded by points XZP_1. Likewise, producer welfare (producer surplus) is the triangular area bounded by points XP_1U.

But consider what happens in a world of completely free trade, in which domestically produced goods have to compete with cheaper imports. In Figure 14.3, imports are priced at the ruling world price of P_W, which is lower than P_1. In this situation, equilibrium now occurs in the domestic market at point V. Although domestic demand has increased to Q_{D1}, domestic supply (located where the domestic supply curve cuts the horizontal price line at P_W) falls to Q_{S1}. Imports (equal to $Q_{D1} - Q_{S1}$) fill the gap between domestic demand and supply.

To understand how imports affect economic welfare within the country, it is important to understand how consumer surplus and producer surplus change after the price of the good falls to the world price P_W. Consumer surplus increases by the wedge-shaped area bounded by the points P_WVXP_1. This divides into two parts, shown on the diagram by areas B and C. Look closely at area B. This shows a welfare transfer away from domestic firms to domestic consumers. The fall in the price from P_1 to P_W, brought about by lower import prices, means that part of the producer surplus that domestic firms previously enjoyed now becomes consumer surplus. The consumers 'win' and the domestic producers 'lose'.

But this is not the end of the story. Consumers enjoy a further increase in consumer surplus, which is brought about by receipt of area C. As a result, the total increase in consumer surplus exceeds the size of the welfare transfer from producer surplus to consumer surplus. A net welfare gain thus results, equal to area C, which makes up part of the consumer surplus that households now enjoy.

SYNOPTIC LINK

Chapter 5 on perfect competition, imperfectly competitive markets and oligopoly explains the key welfare concepts of consumer surplus and producer surplus.

STUDY TIP

You should practise drawing and explaining Figure 14.3 and the diagrams that follow to show the welfare gains and losses resulting from trade and import controls. Remember, however, the assumptions that underpin free-trade theory which we explained earlier. Questioning these assumptions can weaken the case for free trade.

How a tariff or import duty affects economic welfare

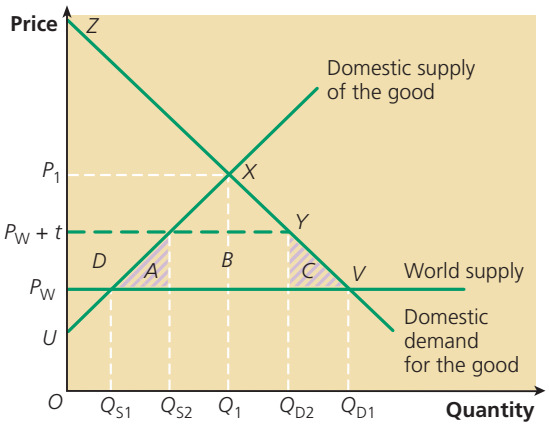


Figure 14.4 The effect of imposing a tariff

We will now assume that domestic firms pressure the government to introduce a tariff to protect the home market. If the tariff equals the distance between P_1 and P_W, the domestic market for the good reverts to the original equilibrium position that existed before imports entered the country. But suppose the government imposes a smaller tariff, which is just sufficient to raise the price of imports (also allowing domestic producers to charge higher prices) to $P_W + t$ in Figure 14.4. At price $P_W + t$, domestic demand falls to Q_{D2}, while domestic supply rises to Q_{S2}. Imports fall from $Q_{D1} - Q_{S1}$ to $Q_{D2} - Q_{S2}$.

At the higher price, consumer surplus *falls* by the wedge-shaped area $P_W VYP_W + t$, which equals the areas $D + A + B + C$. The higher price increases producer surplus by the area D, and the government gains tariff revenue shown by the area B. (The government's tariff revenue is measured by total imports, $Q_{D2} - Q_{S2}$, multiplied by the tariff per unit of imports, $P_W + t - P_W$.) The areas D and B are *transfers* of welfare away from consumers to domestic producers and the government respectively. The *net welfare loss* resulting from the tariff, which is the sum of triangles A and C in the diagram, is:

$$(D + A + B + C) - (D + B), \text{ which equals } A + C$$

Changing comparative advantage and the pattern of world trade

To many people living in industrial countries during the nineteenth century and the first half of the twentieth century, it must have seemed almost natural that the earliest countries to industrialise, such as the UK, had done so because they possessed a competitive and comparative advantage in manufacturing. It probably seemed equally natural that a pattern of world trade should have developed in which the industrialised countries in what is now called the Global North exported manufactured goods in exchange for foodstuffs and raw materials produced by countries whose comparative advantage lay in the production of primary products — that is, the countries in the Global South.

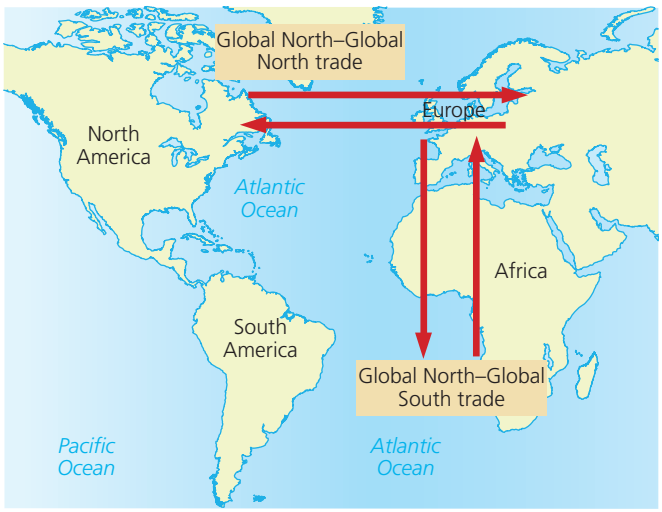


Figure 14.5 Global North–Global North and Global North–Global South patterns of trade

However, in recent years, the pattern of world trade has become quite different from the Global North–Global South exchange of manufactured goods for primary products that characterised the nineteenth century. In a Global North–Global North pattern of trade, which is illustrated in Figure 14.5, the developed industrial economies now exchange goods and services mostly with each other. However, a growing fraction of their trade, particularly in the case of imports, is with newly industrialising countries (NICs) or emerging markets, particularly India, China and South Korea. A group of countries known

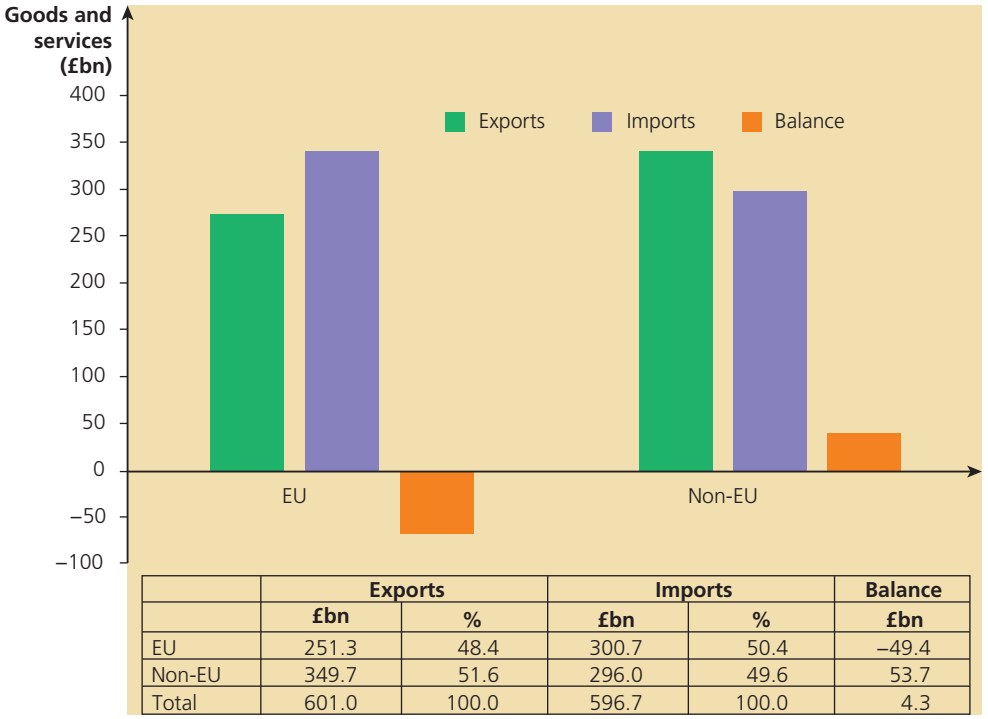
as the BRIC countries (Brazil, Russia, India and China) are responsible for exporting large quantities of goods and services to the Global North. India, China and South Korea now export manufactured goods to countries in the Global North such as the UK and the USA, and import raw materials or commodities such as copper from developing economies such as Zambia. They also import a growing fraction of the crude oil produced by oil-exporting developing economies such as Saudi Arabia and Venezuela.

The shift of manufacturing to China and other NICs reflects changing competitive and comparative advantage and the deindustrialisation of the UK, North America and some of the major European economies. Only a relatively small proportion of the trade of Global North countries is with poorer countries in the non-oil-producing developing world.

TEST YOURSELF 14.4
What is the difference between comparative advantage and competitive advantage?

The pattern of the UK's international trade

The EU has been a significant trading partner of the UK. In 2020, the UK exported £251 billion of goods and services to EU member states (see Figure 14.6). This is equivalent to 48.4% of total UK exports. Goods and services imported from the EU were worth £301 billion (50.4% of the total) in 2020.

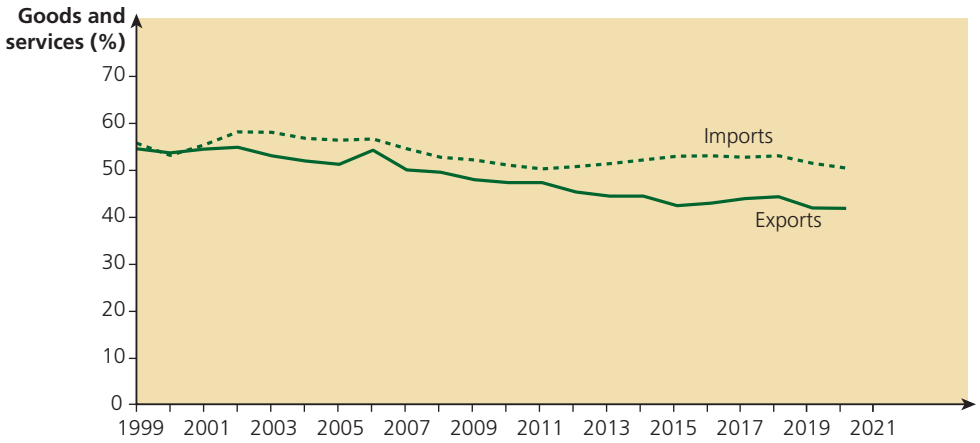


	Exports		Imports		Balance
	£bn	%	£bn	%	£bn
EU	251.3	48.4	300.7	50.4	–49.4
Non-EU	349.7	51.6	296.0	49.6	53.7
Total	601.0	100.0	596.7	100.0	4.3

Source: House of Commons Library

Figure 14.6 UK trade with EU and non-EU countries, 2020

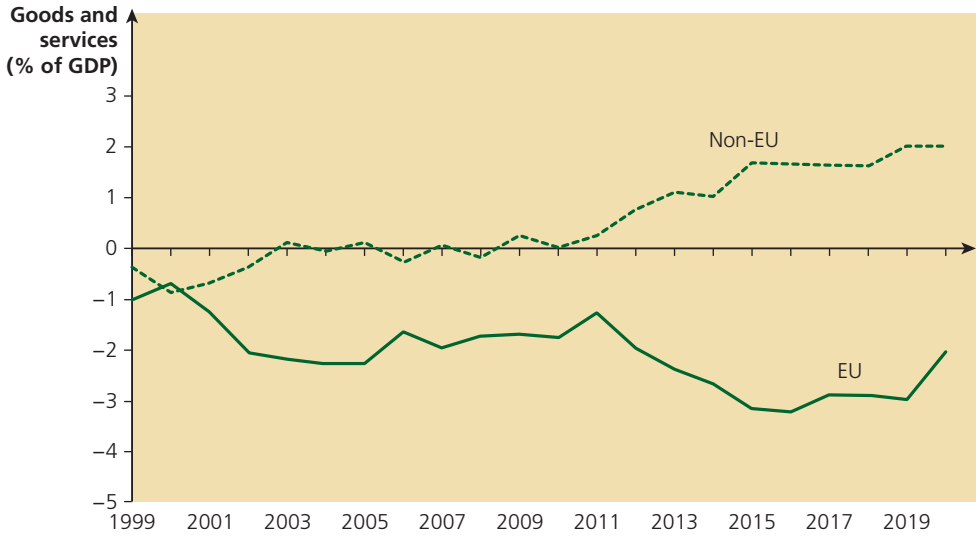
The UK had a trade deficit of –£49.4 billion with the EU in 2020 but a surplus of £53.7 billion with non-EU countries.



Source: House of Commons Library

Figure 14.7 Share of UK trade with the EU, 1999–2020

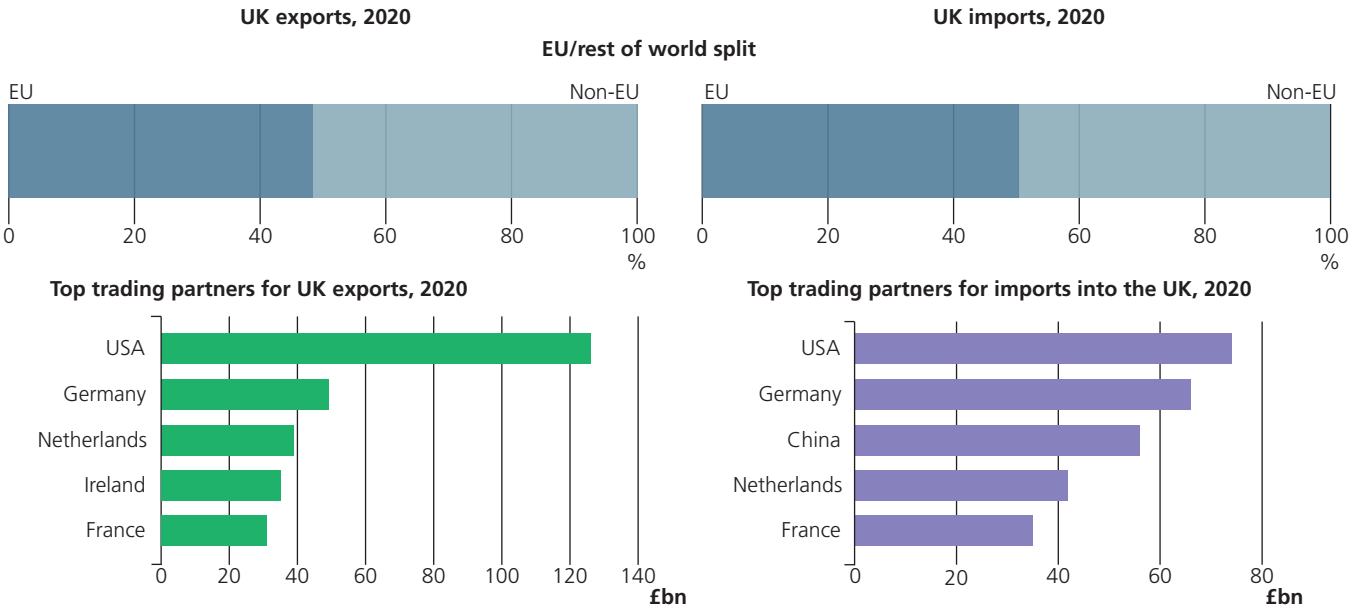
The share of UK exports going to the EU has declined gradually over recent years (see Figure 14.7). In 2002, the EU accounted for almost 55% of UK exports. By 2020, this had fallen to just under 42%. The picture on imports is slightly less clear. In 2002, 58% of UK imports were from the EU. This has fallen since but has remained over 50% over this period. The UK has had a trade deficit with the EU in every year since 1999, peaking at 3.7% of GDP in 2016 before falling to 2.3% in 2020 (see Figure 14.8).



Source: ONS

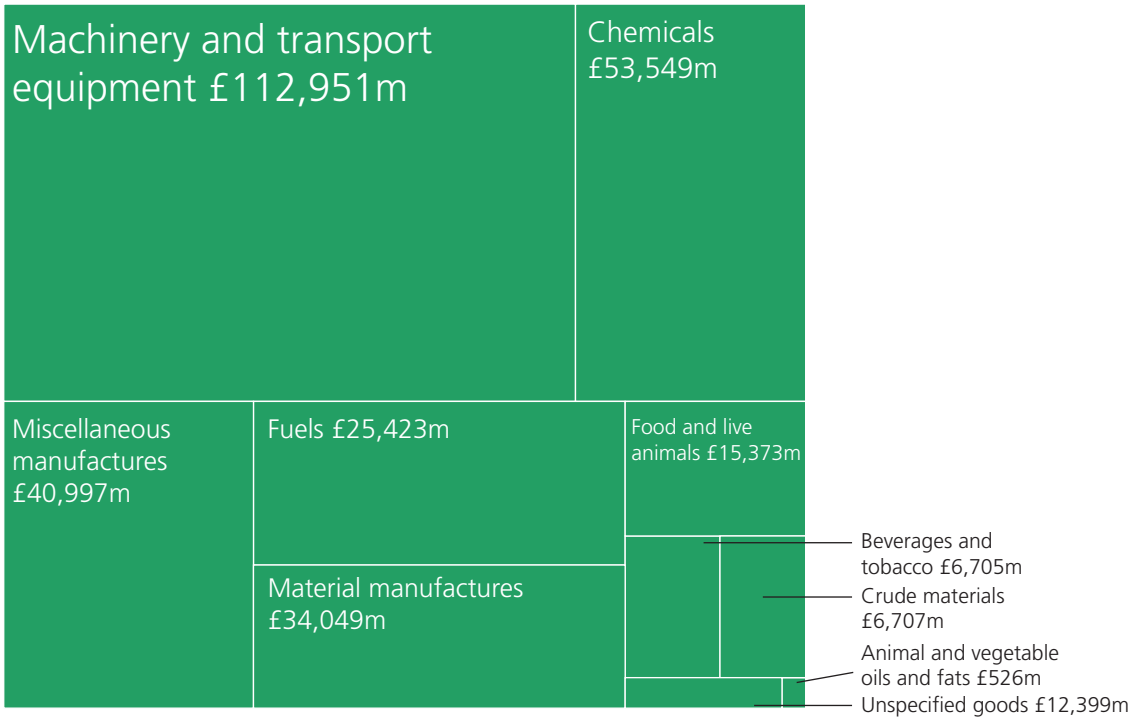
Figure 14.8 Balance of UK trade with EU and non-EU countries, 1999–2020

Figure 14.9 shows that four out of the top five of countries to which the UK exported in 2020 were EU member states, though the USA was by far the main destination for UK exports, leading Germany by over £75,000 million. On the import side, the USA was the UK's main source of imports but the gap between the USA and other nations was smaller than for exports.



Source: ONS

Figure 14.9 The UK's top trading partners for exports and imports, 2020

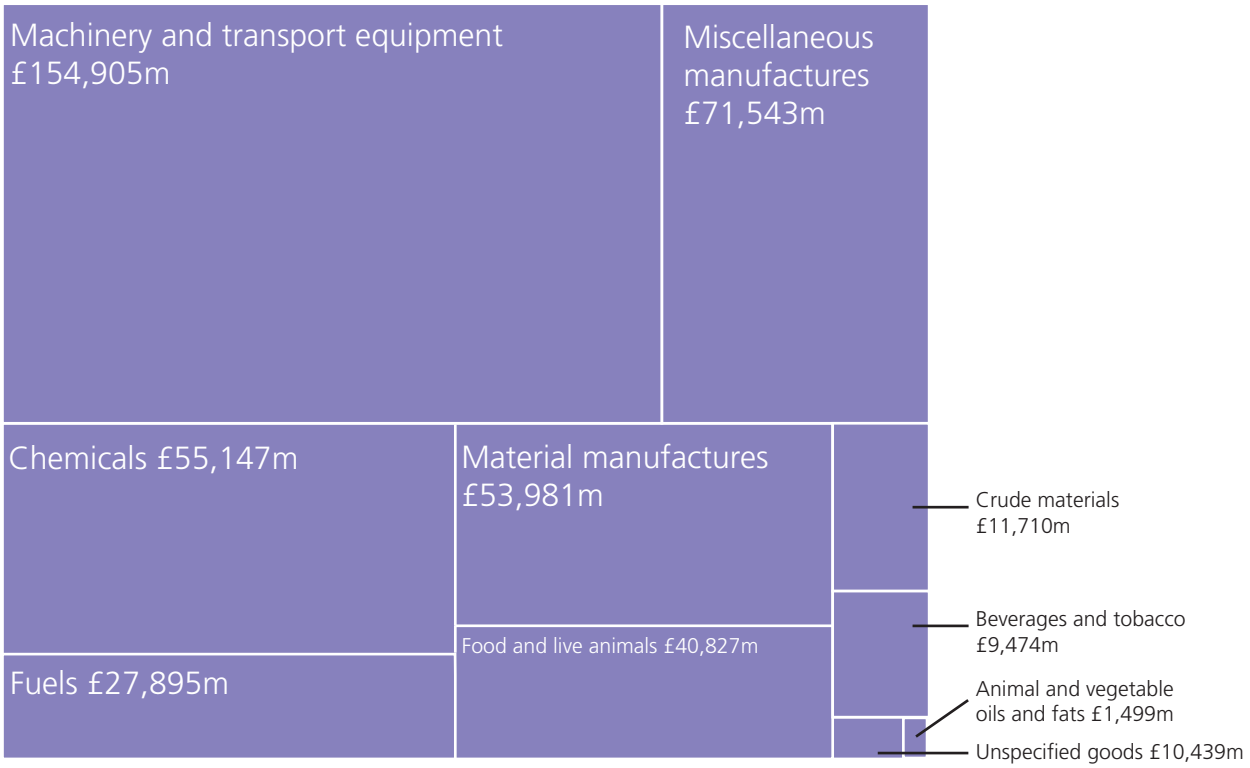


Source: ONS

Figure 14.10 The commodity breakdown for UK exports of goods, 2020 (total = £308,679m)

Finally, Figures 14.10 and 14.11 show the commodity breakdown in the pattern of UK trade in 2018 for exports and imports of goods. The diagrams show the pattern of UK trade in terms of the different groups of goods which are traded.

Machinery and transport equipment headed the list for both exports and imports, with chemicals and miscellaneous manufactures in second place respectively for exports and imports.



Source: ONS

Figure 14.11 The commodity breakdown for UK imports of goods, 2020 (total = £437,420m)

QUANTITATIVE SKILLS 14.1

Worked example: interpreting export and import statistics

Making use of the data in Figures 14.10 and 14.11, calculate the UK's trade deficit or surplus in 2020 for:

(a) machinery and transport equipment
(b) chemicals
(c) fuels

(a) £112,951m – £154,905m = deficit of £41,954m
(b) £53,549m – £55,147m = deficit of £1,598m
(c) £25,423m – £27,895m = deficit of £2,472m

STUDY TIP

Figures 14.6 to 14.8 relate to UK trade in goods and services, whereas Figures 14.10 and 14.11 are for goods only, excluding services. In 2020, as in other years, the UK's exports of services to EU member states were significantly larger than its exports of goods. Make sure you are aware of this, and also are able to interpret comparative statistics.

Degrees of international economic integration

KEY TERMS

free trade area in a free trade area, member countries abolish tariffs on mutual trade, but each partner determines its own tariffs on trade with non-member countries.

customs unions trading blocs in which member countries enjoy internal free trade in goods and possibly services, with all the member countries protected by a common external tariff barrier.

The main types of international economic integration are as follows:

- **Preference areas.** Countries agree to levy reduced, or preferential, tariffs on certain trade: for example, the EU's Association Agreements with certain developing economies.
- **Free trade areas.** Member countries abolish tariffs on mutual trade, but each partner determines its own tariffs on trade with non-member countries. A potential problem here is that traders try to import goods into the partner with the lowest external tariff and then re-export them to others tariff-free from there. To avoid this, complex rules of origin and border controls tend to govern intra-area trade to ensure that free trade refers only to the partners' produce and not to their imports.
- **Customs unions.** There are two main features of a customs union: first, a free trade area where member countries agree to abolish tariffs, quotas and other barriers to trade; second, a common external tariff wall set up on trade with non-member countries and coordinated centrally.
- **Common markets.** These are customs unions with additional provisions to encourage trade and integration through the free mobility of factors of production and the harmonisation of trading standards and practices.
- **Economic unions.** These add further harmonisation in the areas of general economic, legal and social policies, and the development of union-wide policies. Economic union may be supplemented by monetary union, which entails a common currency and monetary policy.
- **Political unions.** These are the ultimate form of economic integration. They involve the submersion of separate national institutions. Even within this category, however, several degrees of integration exist. For example, in the USA the states have substantial powers of taxation, whereas at the moment UK local government has few such powers (though Scotland has some ability to raise its own taxes, such as being able to vary the rates of income tax from those levied in other parts of the UK).

TEST YOURSELF 14.5

Which of the following international organisations is protected by a common external tariff?

A Free trade area
B Customs union
C Preference area
D Monetary union

Free trade areas versus customs unions

Figure 14.12 illustrates the main difference between a customs union and a free trade area, and also the way in which the two forms of trading bloc are similar. The similarity lies in the fact that both organisations have internal free trade, at least in goods though not necessarily in services. The difference lies in the way tariffs are set against imports from non-member states. Members of a free trade area are free to set their own tariffs against non-members, but members of a customs union lose this freedom and have to abide by an external tariff common to all member states.

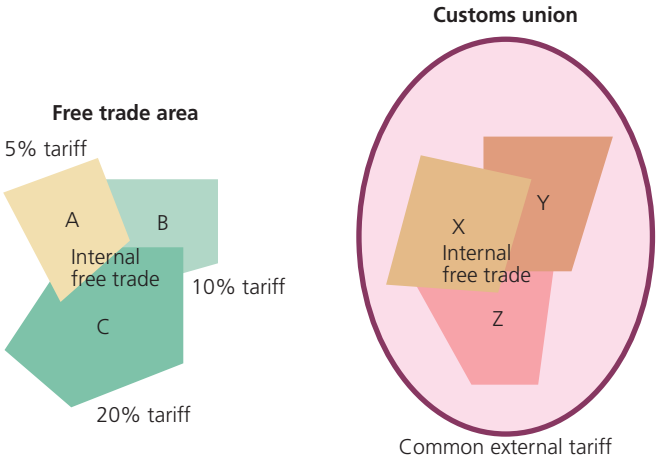


Figure 14.12 The difference between a free trade area and a customs union

EXTENSION MATERIAL

The European Union as an example of international economic integration

The organisation which became the EU in 1993 was originally called the European Economic Community (EEC). The reasons why it was formed almost 60 years ago were partly economic and partly political. The economic reasons for the formation of the EEC stemmed from the benefits that all the member countries believed they could achieve from creating a large area of internal free trade.

On the political front, West Germany (as it was then called), which was one of the original six members of the new organisation, hoped that EEC membership would help to protect Germans from the military threat of invasion by the Soviet Union. By contrast, France hoped that a stable and economically successful Germany would reduce the chance of another major war in Europe. This could best be achieved if EEC member states traded freely with each other.

On the economic front, France was willing to create free trade in manufactured goods (which would mostly benefit Germany), provided small French farms could be protected by a common external tariff from the competition of more efficient New World farmers in countries outside the EEC. Germany readily agreed to the creation of a protectionist common agricultural policy (CAP) as the *quid pro quo* of free trade in manufactured goods within the EEC. (*Quid pro quo* means doing a favour in return for something.)

The EU as an economic union

In the early 1970s, the United Kingdom's decision to join the EEC meant leaving a free trade area and joining a customs union which already possessed a number of common economic policies. Indeed, the fact that the word 'economic' is central to the European Union's original name clearly indicates that the union's founders intended to develop the economic nature of the organisation well beyond a mere trading bloc.

Since the creation of the EEC nearly 60 years ago, the EU has introduced a number of common economic policies. Of these, the common agricultural policy, or CAP, which began in 1962 is by far the most important (and expensive). Other common policies include the common fisheries policy, the EU competition policy and the EU regional policy. There was also an early attempt to introduce a common tax policy, centring on the introduction of VAT as the EEC's main expenditure tax. But member countries continue to levy their own national rates of VAT and tax harmonisation has not been developed much further.

KEY TERM
eurozone (also known as the **euro area**) the name used for the group of EU countries that have replaced their national currencies with the euro. As of 2022, 19 of the 27 EU members have replaced their national currencies with the euro.

Nonetheless, the creation in the late 1990s of the Stability and Growth Pact to constrain or limit government spending and borrowing in **eurozone** countries was an attempt to impose some elements of a common fiscal policy. In very recent years, the European Commission (the EU's executive body) and the European Central Bank (ECB) have imposed stiff conditions on countries in the eurozone, particularly Greece, when lending to them to bail out the huge national debts these countries incurred during the so-called eurozone crisis.

It is believed that the move towards closer economic integration and the freedom of movement of workers across the EU were two of the main reasons why the UK voted to leave the EU in 2016.

TEST YOURSELF 14.6

Explain why the correct answer to the following question is D.

Which trading bloc has the highest level of economic integration?

A United States-Mexico-Canada Agreement

B European Union

C Eurasian Customs Union

D Eurozone

The World Trade Organization

To understand how the World Trade Organization (WTO) came into existence and its role today, it is useful to go back to events occurring in the 1930s and 1940s. In the 1940s, during the Second World War, it was widely believed (especially in the UK and the USA) that the worldwide Depression and mass unemployment of the 1930s had been made worse, and were possibly caused, by a collapse of international trade. Protectionist policies, introduced by countries desperately trying to save local jobs, were blamed.

By 1945 the USA and the UK had decided to try to create a postwar world of free trade. Because this required international agreement, the General Agreement on Tariffs and Trade (GATT) was established as a multilateral agreement to try to liberalise world trade. To begin with, GATT was supposed to be a temporary organisation, to be replaced with a 'world trade organization' as soon as member countries could agree. However, member countries were unable to agree, so the 'temporary' organisation lasted much longer than intended. Indeed, GATT still exists today, though its functions have largely been taken over by the WTO.

Throughout their history, GATT and later the WTO organised rounds of talks among member countries to reduce import controls. The rounds, which took place at roughly five-year intervals, were often named after the city or country in which the talks were initiated: for example, the Tokyo Round, the Uruguay Round and, more recently, the Doha Round. Out of respect, the Kennedy Round in the mid-1960s was named after the then recently assassinated American president. Each round of talks ends with an agreement to reduce import controls. GATT, and latterly the WTO, have then tried to get member countries to implement the tariff cuts they have agreed.

GATT and WTO agreements have been successful in reducing import controls on manufactured goods. There has been much less success in securing agreement to reduce tariffs and quotas on trade in services and agricultural goods. Recently, the WTO has tried to persuade the developed economies of the EU and the USA to open their markets to cheap food imports from the developing world. However, the most recent rounds of talks organised by the WTO at Cancún (in Mexico) and Doha (in Qatar) were not successful. Economists and politicians in many developing economies claim that this lack of success provides further evidence of globalisation and international organisations serving the interests of rich countries at the expense of the poor.

SECTION 14.2 SUMMARY

- International trade widens a country's production possibilities and also its consumption possibilities, and enables countries to benefit from specialisation and economies of scale.
- Absolute advantage means that a country is technically more efficient in producing a good than other countries — it can produce more of the product with the same amount of resources.
- A country has a comparative advantage in a product when its opportunity cost of production is lower than in another country.
- According to the principle of comparative advantage, specialisation and trade can lead to an increase in world output, which can translate into a net welfare gain. The theory shows that, provided each country has a comparative advantage in the production of a good, each country can benefit from specialisation and trade.
- Increased international competition and the potential for economies of scale can result in additional gains from trade.
- Import controls and other forms of protectionism have been justified by strategic trade theory, the protection of infant industries and a variety of other economic and non-economic arguments.
- The Global North–Global South pattern of world trade has largely given way to a Global North–Global North pattern.
- The European Union (EU) started life in the 1950s as the EEC, a customs union, but since 1993 it has been a much fuller economic union.
- Much of the UK's trade is with EU member states.

14.3 The balance of payments

KEY TERM
balance of payments a record of all the currency flows into and out of a country in a particular time period.

The **balance of payments** is the record of all money flows or transactions between the residents of a country and the rest of the world in a given period of time. These transactions can be broken down into three main accounts:

- the current account
- the capital account
- the financial account.

An overview of the UK balance of payments in 2020 is shown in Table 14.9.

Table 14.9 Selected items from the UK balance of payments (£m), 2020

	£m
The current account (mostly trade flows)	
Balance of trade in goods	–128,741
Balance of trade in services	133,021
Primary income flows *(Net income flows)*	–31,996
Secondary income flows *(Net current transfers)*	–28,172
Balance of payments on the current account	**–55,888**
The capital account (Transfers, which used to be in the current account)	**–2,610**
The financial account *(Capital flows, which used to be in the capital account)* Net direct investment	–65,174
Net portfolio investment	–3,357

	£m
Financial derivatives and employee stock options	30,383
Other capital flows *(including many short-term 'hot money' flows)*	–36,357
Drawings on reserves	–2,582
Net financial transactions *(This is the overall balance for the financial account.)*	**–77,087**
Net errors and omissions *(The financial account balance minus the current and capital account balances. It is required to compensate for inevitable mistakes and ensure that these three accounts sum to zero.)*	**–18,589**

Note: the ONS has changed how it presents figures in the financial account. Figures that used to be shown as a plus are now shown as a minus and vice versa. Consequently, mathematically, the balances for the account as a whole no longer sum to zero.

Source: ONS, UK Balance of Payments, *The Pink Book*, 2021

The difference between the current, capital and financial accounts of the balance of payments

The current account of the balance of payments

KEY TERM
current account measures all the currency flows into and out of a country in a particular time period in payment for exports and imports of goods and services, together with primary and secondary income flows (previously known as income flows and transfers).

The balance of payments on the **current account** measures the flow of expenditure on goods and services, and shows the difference between the amount received from exports and the amount paid for imports. The current account is usually regarded as the most important part of the balance of payments because it reflects the economy's international competitiveness and the extent to which the country may or may not be living within its means. The current account is the main area of study at A-level and will be examined in detail in the following sections of this chapter.

TEST YOURSELF 14.7
Explain why the correct answer to the following question is D.

The current account of the balance of payments comprises:

A cash deposits held by high street banks with the central bank

B government revenues from taxation minus its planned spending

C all transactions involving money leaving or entering the country

D trade in goods and services, primary income and secondary income

The capital account of the balance of payments

Several years ago, a change was made to the way the UK balance of payments is presented. The items in the capital account, a long-standing part of the balance of payments, were changed. Capital flows were moved out of the capital account and listed under a new heading: the financial account. The capital account now comprises various transfers of income that were part of the current account before the new method of classification was adopted. It is now a very small, and usually insignificant, part of the overall balance of payments account.

KEY TERM
financial account the part of the balance of payments which records capital flows into and out of the economy.

The financial account of the balance of payments

The **financial account** of the balance of payments has five main components which you can see towards the bottom of Table 14.9. It is worth noting that, if a country is living beyond its means by running a current account deficit, the borrowing which finances the deficit is recorded in either the capital account or the financial account.

STUDY TIP
A detailed knowledge of the financial account is not needed at A-level, but we will take a look at capital flows later in this section and explain the difference between foreign direct investment (FDI) and portfolio investment.

The UK's balance of payments on current account

The four components of the current account

The main sections of the current account, which are shown in Table 14.9, are the balance of trade in goods, the balance of trade in services, **primary income** flows and **secondary income** flows. (These are summarised in Table 14.10.) As we have already noted, the current account is usually regarded as the most important part of the balance of payments because it reflects an economy's international competitiveness and the extent to which the country is living within its means. If the currency outflows in the current account exceed the currency inflows, there is a **current account deficit**. If receipts exceed payments, there is a **current account surplus**.

KEY TERMS
balance of primary income comprises inward income flowing into the economy in the current year, generated by UK-owned capital assets located overseas (e.g. shares owned in foreign companies, or land owned overseas), and outward flows of income from the economy in the current year, generated by overseas-owned capital assets located in the UK.

balance of secondary income current transfers (e.g. gifts of money, international aid and transfers between the UK and the EU), flowing into or out of the UK economy in a particular year.

current account deficit when currency outflows in the current account exceed the currency inflows.

current account surplus when currency inflows in the current account exceed the currency outflows.

Some older textbooks may still refer to the terminology used in the balance of payments prior to 2014. Primary income was previously referred to as 'investment income' and secondary income was previously known as 'net transfers'.

Table 14.10 Sections of the current account, 2020 (£m)

Balance of trade in goods	–128,741
Balance of trade in services	133,021
Balance of primary income	–31,996
Balance of secondary income	–28,172
Balance of payments on the current account	**–55,888**

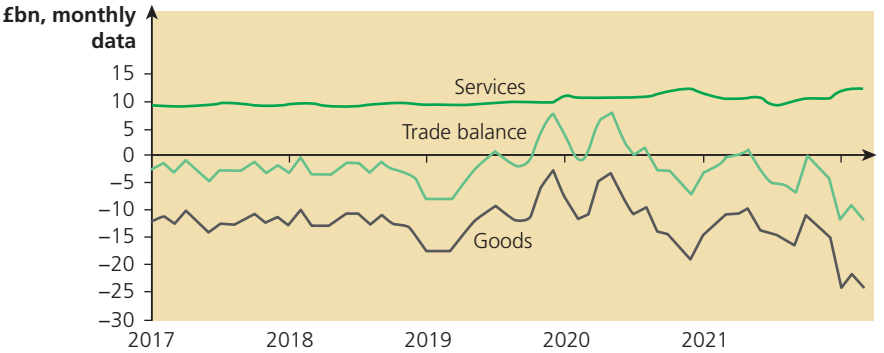
Source: ONS, UK Balance of Payments, *The Pink Book*, 2021

STUDY TIP

Make sure you understand the different items in the current account relating both to trade and to primary and secondary income flows. You may need to calculate the trade balance and this would only include trade flows.

For each of the sections of the current account, and for the current account itself shown in the bottom row of Table 14.10, a plus sign (+) indicates a credit item (net currency flowing into the UK), and a minus sign (–) indicates a debit item (net currency flowing out of the UK). The balance of trade in goods is sometimes called the balance of *visible* trade, and the balance of trade in services is part of the balance of *invisible* trade.

The story that can be read into the figures is quite worrying. Look carefully at Figure 14.13. Throughout the period shown the UK has run a persistent deficit on the current account. The value of sterling has depreciated significantly since 2008 but the UK has continued to run a large deficit. In 2016 the deficit on the current account was 5.3% of GDP, the largest of the G7 countries. The second largest was Canada at 3.1%. In the words of Mark Carney, the then governor of the Bank of England, the UK is reliant on the 'kindness of strangers' to finance the deficit. Put simply, the UK is living beyond its means and relies on investment flows from abroad which are secured against assets such as office buildings which are located in the UK. By 2020, the UK current account deficit had narrowed to 2.6% of GDP — similar (as a proportion of their respective GDPs) to the USA, France and Canada.



Source: House of Commons Library

Figure 14.13 Changes in the UK trade balance, 2017–21

As Figure 14.13 shows, the UK balance of trade has been in deficit but relatively stable in recent years, with a goods deficit being partially offset by a services surplus. The current account deficit, which is not shown in

Figure 14.13, has widened because the primary income deficit has been growing. This means that foreigners who own assets in the UK are earning more from their assets than UK investors are receiving from their investments overseas. The primary income deficit has been the driving force behind the widening current account deficit. In earlier years, the UK enjoyed a substantial surplus of investment income.

Notice that the balance of trade actually moved into surplus in 2020 (for the first time since 1997) largely as a result of the slowdown in trade due to the Covid-19 pandemic. It is possible but unlikely that the surplus on the balance of trade will continue in the near future.

The consensus view among economists is that the size of the current account deficit is unlikely to cause a crisis in the near term, but is storing up problems for the future. If international investors lose confidence in the financial assets sold by the UK, the country will experience a dramatic fall in living standards because it will be unable to finance its trade deficit.

QUANTITATIVE SKILLS 14.2

Worked example: interpreting changes in the main sections in the current account

Table 14.11 shows selected components of a country's balance of payments on current account as a percentage of GDP in 2022 and 2023.

Table 14.11 Balance of payments: selected components, 2022–23 (% of GDP)

Year	Balance of trade in goods and services	Primary income balance	Secondary income balance	Current account balance
2022	–4.0	+2.0	–1.4	–9.0
2023	–3.9	–4.6	–1.3	–10.2

Do the data suggest that a main reason for the rise in the current account deficit as a percentage of GDP between 2022 and 2023 was a significant deterioration in the primary income balance?

The answer is 'yes'. Over the two-year period, the current account deficit increased by 1.2% of GDP. Of the three components of the current account shown in the table, the balance of trade in goods and services deteriorated by 0.1% of GDP, while the secondary income balance improved by 0.1% of GDP. However, the primary income balance deteriorated by the significantly larger percentage of 6.6%, from +2.0% of GDP to –4.6%. This suggests that the deterioration in the current account balance was mainly caused by the deterioration in the primary income balance, though we cannot be absolutely sure because we don't know the sizes of all the items in the table.

KEY TERM

balance of trade in goods the part of the current account measuring payments for exports and imports of goods. The difference between the total value of exports and the total value of imports is sometimes called the 'balance of visible trade'.

The main sections of the current account

We will now take a closer look at each of the four sections of the current account.

The balance of trade in goods

The balance of trade in goods shows the extent to which the value of exports of goods exceeds the value of imports, or vice versa. Table 14.12 does not indicate the total values of exports and imports, which were respectively £308,679 million and £437,420 million in 2020. The balance of trade in goods was therefore in deficit to the tune of £128,741 million. This figure is shown in the bottom row of Table 14.12.

The balance of trade in goods can also be disaggregated (broken up) into different forms of trade in goods, such as the balances of trade in

manufactured goods and non-manufactured goods. Some of the different ways of disaggregating the balance of trade in goods are shown in Table 14.12.

Table 14.12 Selected items from the UK balance of trade in goods, 2020 (£bn)

Balance of trade in food, drinks and tobacco	–28.2
Balance of trade in raw materials (basic goods)	–6.0
Balance of trade in oil	+2.0
Balance of trade in manufactured goods	–94.0
Balance of trade in all goods	**–128.7**

Source: ONS, UK Balance of Payments, *The Pink Book*, 2021

STUDY TIP

You must avoid confusing the balance of trade in goods with the whole of the current account, and the current account with the balance of payments as a whole. The balance of payments includes the current account but includes other parts of foreign flows of money as well.

TEST YOURSELF 14.8

Explain why the correct answer to the following question is B.

All other things being equal, an appreciation in the value of sterling is likely to:

A increase the price of imported goods

B widen the size of the UK's deficit on the current account

C reduce the size of the UK's budget deficit

D make UK exports more price competitive on world markets

EXTENSION MATERIAL

Data presentation

Take note of the word 'Selected' in the captions of Tables 14.12 and 14.13. The word tells you that not all the items in the balance of trade in goods and the balance of trade in services are in the tables. Selections such as these try to separate the most interesting items from the 'background noise' of other, perhaps less interesting items.

Table 14.12 shows that the UK is a net importer of primary products (food and raw materials). The surplus on the balance of trade in oil was largely the result of the sharp fall in demand for oil as travel plans were severely restricted in 2020 due to the extended 'lockdown' of the UK economy due to the Covid-19 pandemic.

The balance of payments deficit in manufactured goods is significant. Apart from the periods during and immediately following the First and Second World Wars, for over 200 years Britain was a net exporter of manufactured goods. In the mid-nineteenth century, it was the 'workshop of the world'. This has now changed. In the early 1980s, the UK became a net importer of manufactured goods. The manufactured goods deficit is now huge, reflecting loss of competitiveness, the resulting deindustrialisation of the UK, and the fact that most manufactured goods are now produced in emerging-market countries, particularly in China.

STUDY TIP

Make sure you have some knowledge of the balances of trade in items such as manufactured goods, oil and automobiles. Knowledge of what the UK actually produces can be useful to obtain application marks in answers.

The balance of trade in services

The decline of manufacturing and the growth of service industries mean that the UK now has a post-industrial and service-sector economy. This is reflected in the **balance of trade in services** shown in Table 14.13.

KEY TERM

balance of trade in services part of the current account, the difference between the payments for the exports of services and the payments for the imports of services.

Table 14.13 Selected items from the UK balance of trade in services, 2020 (£bn)

Balance of trade in transport	+6.2
Balance of trade in travel	–1.7
Balance of trade in telecommunication, computer and information services	+15.9
Balance of trade in insurance and pension services	+17.9
Balance of trade in financial services	+45.8
Balance of trade in intellectual property	+4.9
Balance of trade in all services	**+133.0**

Source: ONS, UK Balance of Payments, *The Pink Book*, 2021

Whereas most manufactured goods are internationally tradable, the same is not generally true for services such as retailing, car repair and hairdressing. Services like these are produced and consumed in the non-internationally traded economy, or sheltered economy. However, many services which were previously produced within the UK are now being imported. This is an important part of the globalisation process.

UK companies, which used to produce many services 'in-house', now outsource or buy in the services from outside suppliers, often located in countries with cheap labour. UK-based companies are locating 'back-office' service activities overseas, including many financial and ICT-related services. Many call centres providing customer services and direct marketing services have moved to India, although some are returning to the UK, using automated systems to reduce costs.

Nevertheless, as Table 14.13 shows, the UK is still a significant net exporter of financial, insurance and ICT services. These industries illustrate the UK's competitive advantage in service-sector industries, though the picture is not as favourable in industries such as travel and tourism, where Britons now spend much more in other countries than overseas residents spend in the UK.

STUDY TIP

Make sure you appreciate the importance of financial services in the balance of trade in services. You should also appreciate the significance of exports of financial services to the UK economy. Although the current account and balance of trade are usually in deficit, the surplus on trade in services is largely derived from the financial services industry.

APPLICATION OF ECONOMICS IN THE REAL WORLD 14.4

The UK financial services industry

We wrote this case study in the aftermath of the 2008–09 recession. Since then, with recovery under way, the future of the UK financial services industry may be more promising than was thought in 2010 and 2011 when the UK economy was 'flat-lining'. However, the UK's exit from the European Union may change this.

The financial services sector is a significant contributor to UK income and employment. Over a million people in the UK are employed in financial services, of which two-thirds are based outside London. The financial services sector is also one of the largest export industries in the UK. However, on the negative side, the growth of the financial services industry has led to an unbalanced economy.

The financial crisis that began in the summer of 2007 has, however, highlighted the need for a more resilient and sustainable financial services industry to support the broader economy. In addition to being an important part of the economy in its own right, the financial services sector provides essential credit and financial services to businesses and households.

During the period after 2000 the expansion of the financial sector was a significant influence on the growth rate of the economy overall. But the recent growth of the sector has been reversed, reducing GDP permanently by about 1.9%. The country will suffer a further loss of income as a result of the losses that banks have made, giving a total fall in national income of about 2.4%, and reducing government revenue by about 1% of GDP.



Financial services are one of the UK's largest exports

Follow-up questions

1. Explain two reasons for the temporary decline of the UK financial services industry after 2007.
2. 'The growth of the financial services industry has led to an unbalanced economy.' What does this mean and how might the UK economy be 'rebalanced'?
3. Whereas many service industries are sheltered from international trade, this is generally not true for financial services. Explain why this is so.
4. Now that the UK has left the European Union, describe how its exit has affected the UK's financial services industry.

Primary income

Primary income flows are net income flows made up mostly of investment income generated from profits, dividends and interest payments flowing between countries. When we look later at the financial account of the balance of payments, we will explain how UK-based multinational companies (MNCs) invest in capital assets located in other countries. The profit income generated from overseas investment flows back to the parent company and its UK shareholders. The investment itself is an outward capital flow, but the income it generates is current income, figuring in the current account of the balance of payments.

Profits also flow out of the UK to the overseas owners of assets located in the UK — for example, to Japanese or US multinational companies owning subsidiary companies in the UK. In Table 14.10, the item 'primary income flows' is largely determined by the difference between these inward and outward profit flows resulting from capital investment undertaken in the past. The fact that the UK's net income flows (primary income flows)

were –£32 billion in 2020 (up from a net outflow of £19 billion in 2019) indicates that in that year, UK companies owned less profitable assets in the rest of the world than overseas-based MNCs owned in the UK. Until 2011, inward primary income flows had usually exceeded outward flows, as can be seen in Figure 14.20 on page 566. The collapse of foreign investment earnings was a main factor contributing to the current account deficit widening after 2011.

Finally, it is worth remembering that not all primary income flows are profit payments generated by multinational companies on their direct investments abroad. A significant proportion of the income flows are generated from portfolio investment (financial assets such as shares and bonds). Interest and dividend payments within the international financial system contribute significantly to net primary income flows. Make sure you keep up to date with changes in the UK current account.

Secondary income

Secondary income flows are current transfers of income arising from such items as gifts between residents of different countries, donations to charities abroad, and overseas aid. The UK has long had a negative secondary income balance, caused by the UK's net contributions to the EU budget, overseas aid and, in some years in the past, the cost of maintaining armed forces in countries such as Afghanistan.

QUANTITATIVE SKILLS 14.3

Worked example: making calculations from balance of payments data

Table 14.14 shows the balance of payments on current account of a particular country.

Table 14.14 Balance of payment on current account, 2020–23 (£bn)

	2020	2021	2022	2023
Balance of trade in services (£)	+240	+120	+20	+30
Net primary income flows (£)	–30	–35	+10	–15
Net secondary income flows (current transfers) (£)	+5	–2	+12	–8
Current account balance (£)	+200	+100	+80	+100

Calculate the country's balance of trade in goods in each of the four years. Comment on your answers.

The calculations involve subtracting the balance of trade in services, net income (primary income) and current transfers (secondary income) from the current account balance in each of the four years. The answers are:

2020 –£15bn

2021 +£17bn

2022 +£38bn

2023 +£93bn

Over the four-year period, the balance of trade in goods moved from a deficit of £15 billion in 2020 to a surplus of £93 billion. The current account as a whole and the balance of trade in services were in surplus throughout the period. Primary income flows were in deficit throughout the period (except for 2022), which means that more profits and other income were flowing into the country than flowing out. Secondary transfers switched from being positive to negative, and then to positive and negative again, but the secondary transfer flows were generally smaller than the net trade and primary income flows.

TEST YOURSELF 14.9

Figure 14.14 shows a country's trade deficit in goods and its current account balance for the years 2018–23. Suggest one possible reason which might explain why the trade deficit is smaller than the current account deficit.

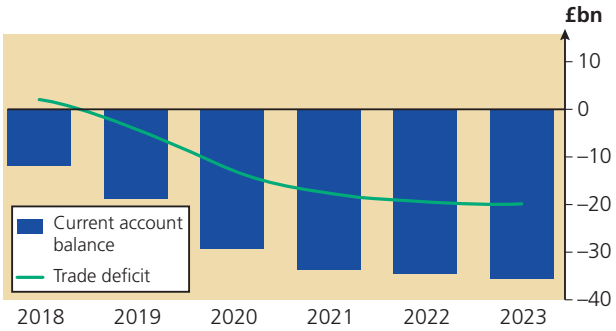


Figure 14.14 The balance of payments on current account and the trade deficit for a particular country, 2018–23

The financial account and investment flows between countries

As our discussion of primary income flows in the current account indicates, it is important to avoid confusing capital flows, which figure in the financial account of the balance of payments, with investment income, which is the predominant share of the primary income flows in the current account. As explained, outward capital flows generate inward flows of investment income in subsequent years. The capital outflow enlarges the stock of capital assets located in other countries, owned by residents and MNCs based in the country exporting the capital. Net capital flows are the difference between inward and outward capital movements. Positive net outward capital flows (i.e. money flowing out of one country to be invested in either physical or financial assets in other countries), over a period of years, mean that the country acquires capital assets located in other countries that are greater in value than the country's own assets bought by overseas companies.

The positive net investment income flow (in the current account) which occurred in years prior to 2012 suggests that, in these years, UK residents and MNCs had invested in a larger and more profitable stock of capital assets in the rest of the world than that acquired by overseas residents and MNCs in the UK. Following the UK's abolition of virtually all foreign exchange controls in 1979, the UK became a large net exporter of capital, presumably because owners of UK MNCs believed that investment abroad would be more profitable than investment in the UK. During the 1980s, the positive net capital outflow meant that the UK became a large owner of overseas capital assets. However, by 2010 the UK had once again become a net debtor nation. In 2016, the largest holders of foreign investment in the UK came from the USA (£3,294bn), France (£878bn), Germany (£858bn), the Netherlands

(£566bn), Ireland (£551bn), Luxembourg (£489bn), Japan (£452bn) and Switzerland (£368bn).

Within the financial account of the balance of payments, we will distinguish between long-term direct capital flows, long-term portfolio capital flows, and short-term speculative 'hot money' capital flows.

Long-term direct capital flows

Direct overseas investment involves acquisition of real productive assets, such as factories, oil refineries, offices and shopping malls, located in other countries. On the one hand, a UK-based MNC may decide to establish a new subsidiary company in the USA, for example. Direct investment can also involve acquisition, through merger or takeover, of an overseas-based company. These are examples of outward direct investment. Conversely, the decisions in the 1980s and 1990s by the Japanese vehicle manufacturers Nissan, Toyota and Honda to invest in automobile factories in the UK led to inward direct investment, or inward **foreign direct investment** (FDI).

KEY TERM
foreign direct investment investment in capital assets (e.g. manufacturing and service industry capacity) in a foreign country by a business with headquarters in another country. Very often the overseas company establishes subsidiary companies in the countries in which it is investing.

Long-term direct capital flows can partly be explained by competitive advantage. The flows are a response to people's decisions to invest in economic activities and industries located in countries that have a competitive advantage. Comparative advantage may also rest in the same country. But since changes in competitive and/or comparative advantage usually take place quite slowly, long-term direct capital flows tend to be relatively stable and predictable.

Portfolio capital flows

These involve the purchase of financial assets (i.e. pieces of paper or, increasingly these days, electronic claims laying claim to the ownership of real assets) rather than physical or directly productive assets. Typically, **portfolio investment** occurs when fund managers employed by financial institutions, such as insurance companies and pension funds, purchase shares issued by overseas companies, or securities issued by foreign governments.

KEY TERM
portfolio investment the purchase of one country's securities (e.g. bonds and shares) by the residents or financial institutions of another country.

The globalisation of world security markets or capital markets and the abolition of exchange controls between virtually all developed economies have made it easy for UK residents to purchase shares or bonds that are listed on overseas capital markets. This has led to a massive increase in portfolio investment. UK residents can now buy shares and corporate bonds that were previously only available on the capital market of the company's country of origin. Securities issued by foreign governments, such as US Treasury bonds, can also be bought.

The credit crunch that began in the USA in 2007, and the so-called 'financial meltdown' that followed, had a significant adverse effect on portfolio investment both within and between countries. Many financial assets, particularly those bought and sold by banks, became known as 'toxic assets'. This term arose from the fact that a potential purchaser of a package of financial assets offered for sale by a bank could not know in advance whether

assets in the package were of high risk and potentially little value, or a sound investment (even the bank trying to make the sale might not know). In such conditions of imperfect information, trading in many types of financial asset collapsed.

Short-term speculative 'hot money' capital flows

Short-term capital movements, which are also called **hot money** flows, are largely speculative. The flows occur because the owners of funds, which include companies and banks as well as wealthy private individuals, believe that a quick speculative profit can be made by moving funds between currencies. Speculating that a currency's exchange rate is about to rise, owners of funds move money into that currency and out of other currencies whose exchange rates are expected to fall. 'Hot money' movements are also triggered by differences in interest rates. Funds flow into currencies with high interest rates and out of currencies with lower interest rates. International crises, such as the outbreak of a war in the Middle East, also cause funds to move into the currency of a 'safe-haven' country, which is regarded as politically stable.

KEY TERM
Short-term speculative capital (also known as **hot money**) refers to flows of money across international borders, moving to find the greatest short-term return; it is usually very easy to move and will move quickly as returns and opportunities arise.

If the pool of 'hot money' shifting between currencies were small, few problems would result. However, short-term capital flows have grown significantly over the last 65 or so years. A large-scale movement of funds from one currency to another creates an excess supply of the former currency and an excess demand for the second currency. To eliminate excess supply and demand, the exchange rates of the two currencies respectively fall and rise. As a result, the movement of funds between currencies produces the changes in exchange rates that speculators (financial traders, in effect, gambling on changes in market conditions so as to maximise their returns) were expecting.

More importantly, a large-scale 'hot money' flow of funds between currencies destabilises exchange rates, the current accounts of balance of payments and, indeed, domestic economies. Such destabilisation occurred late in 2008 and early in 2009 when owners of 'hot money' shifted their funds out of the pound on a massive scale. However, in the next few years, market sentiment changed with the result that 'hot money' flowed into the pound, at least until the middle of 2015. As of 2022, the main 'hot money' flow is into the US dollar, forcing up its value. This is largely due to the dollar being seen as a safe investment in times of uncertainty — caused by the Covid-19 pandemic and Russia's invasion of Ukraine.

Speculative capital flows between currencies such as the dollar, the pound and the euro, which occupy a central place in the finance of international trade, can destabilise the international monetary system. The most recent examples of destabilisation followed the credit crunch and the financial meltdown that were referred to earlier. Banks and other financial institutions, and also governments, in a range of countries (which included the UK) saw their international credit ratings downgraded. To fight the recession that was hitting their economies, governments built up massive budget deficits, which they tried to finance in part by borrowing overseas.

QUANTITATIVE SKILLS 14.4

Worked example: calculating missing statistics in a balance of payments table

Table 14.15 shows some of the items in the balance of payments for a country in 2023.

Table 14.15 Selected items from a balance of payments, 2023

Item	£m
Balance of trade in goods	+1,000
Balance of trade in services	?
Balance of trade in goods and services	**+600**
Primary income	–75
Secondary income	+30
Balance of payments on current account	**?**
Net direct investment	+300
Net portfolio investment	–100
Other items in the financial account	–550
Financial account balance	**?**

Assuming that 'net errors and omissions' and the 'capital account balance' are zero in the balance of payments, fill in the missing numbers in the table indicated by question marks, in each case including a plus or minus sign.

The missing numbers are shown in bold in Table 14.16.

Table 14.16 Selected items from a balance of payments, 2023: completed version

Item	£m
Balance of trade in goods	+1,000
Balance of trade in services	**–400**
Balance of trade in goods and services	+600
Primary income	–75
Secondary income	+30
Balance of payments on current account	**+555**
Net direct investment	+300
Net portfolio investment	–100
Other items in the financial account	–550
Financial account balance	**–350**

TEST YOURSELF 14.10

Table 14.17 shows the components of the balance of payments on current account for an economy in 2023.

Table 14.17 Balance of payments current account, 2023

	£m
Balance of trade in goods	–700
Balance of trade in services	–200
Primary income balance	230
Secondary income balance	–50

Calculate the value of the country's balance of payments on current account.

Applying *AD/AS* analysis to the current account of the balance of payments

In Chapter 10, the meaning of aggregate demand in the economy was explained, together with the aggregate demand equation: $AD = C + I + G + (X - M)$. We showed how, in an *AD/AS* graph, an increase in *any* of the components of aggregate demand (C, I, G or $(X - M)$) causes the aggregate demand (AD) curve to shift rightwards, leading to a new equilibrium national income.

We will now apply *AD/AS* analysis to explain how a change in net exports, or $(X - M)$, affects the national economy. As mentioned earlier, the current account includes non-trade items (investment income and transfers, or primary and secondary income) as well as exports and imports. However, for the rest of this section, we will assume that exports and imports of goods and services are the only two sections of the current account of the balance of payments. Given this simplifying assumption, there is a current account surplus when net exports are positive (i.e. $X > M$), and a current account deficit when net exports are negative (i.e. $X < M$).

Exports are an injection of spending into the circular flow of income, whereas imports are a leakage or withdrawal of spending from the flow (refer to Figure 10.3 in Chapter 10 and to the accompanying explanation of the circular flow diagram).

Suppose initially that $X = M$, which means there is neither a surplus nor a deficit in the current account. Note also that in this situation, given the assumption of no non-trade flows in the current account, foreign trade injections into the circular flow of income exactly equal foreign trade withdrawals from the flow. When $X = M$, the current account has a neutral effect on the state of aggregate demand and on the circular flow of income.

However, at the next stage, overseas demand for UK exports increases, but UK demand for imports remains unchanged. This means there is a net injection of spending into the circular flow of income. The current account moves into surplus, with $X > M$.

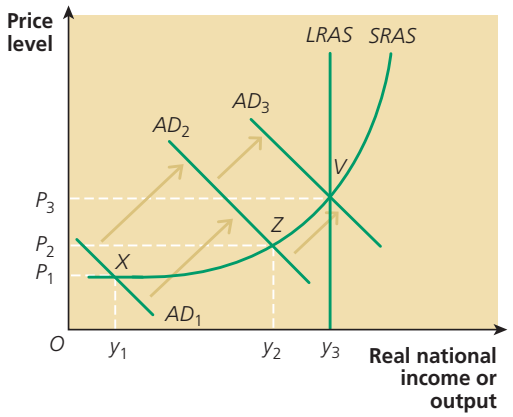


Figure 14.15 How an increase in exports can affect the national economy

In the *AD/AS* diagram in Figure 14.15, the increase in exports shifts the *AD* curve to the right. What happens next in the economy depends on the shape and slope of the *SRAS* curve around the initial point of equilibrium national income. In Figure 14.15, equilibrium national income is initially at point *X*, which shows the economy in deep recession, suffering from deficient aggregate demand. In this situation, any event that increases aggregate demand increases the level of real output in the economy and causes demand-deficient unemployment to fall. An increase in exports is just such an event, shifting the *AD* curve from AD_1 to AD_2. This causes real output to rise from y_1 to y_2, though at the cost of some inflation, since the price level rises from P_1 to P_2.

Following the rightward shift of the aggregate demand curve to AD_2, equilibrium national income is now shown at point *Z*. As the *SRAS* curve becomes steeper, moving up the curve, the diagram tells us that the main effect of further shifts of the *AD* curve fall increasingly on the price level rather than on output and jobs. When the *AD* curve shifts from AD_2 to AD_3 output increases, from y_2 to y_3, but the price level also increases to P_3.

As full employment approaches, export demand becomes *inflationary* rather than *reflationary*.

Nevertheless, in this situation, the growth in export demand eliminates the demand deficiency previously existent in the economy. The economy ends up on its long-run aggregate supply (*LRAS*) curve, with equilibrium national income at point *V*.

At point *V*, what happens next in the economy depends on assumptions made about the nature of short-run and long-run aggregate supply. In Figure 14.15, when the economy produces on the vertical *LRAS* curve, any further increase in the demand for exports leads only to the price level rising above P_3, without any sustained increase in real output. However, there is another possibility. Foreign demand for a country's exports may be a response to favourable supply-side conditions in the domestic economy which shift the *LRAS* curve to the right. This means the economy can produce and supply the goods needed to meet the increase in export demand without generating inflation. This is the desired result of **export-led growth**. The German and Japanese economies enjoyed export-led growth from the 1960s to the 1980s, and China is now enjoying similar benefits. However, the worldwide growth of demand for Chinese exports has begun to cause inflation in the Chinese economy.

KEY TERM

export-led growth in the short run, economic growth resulting from an increase in exports, which is one of the components of aggregate demand; in the long run, economic growth resulting from the growth and increased international competitiveness of exporting industries.

STUDY TIP

You must be able to use the *AD/AS* model and the circular flow of income to analyse how changes in exports and/or imports affect macroeconomic performance (i.e. economic growth, employment and inflation).

Deficits and surpluses on the current account

The current account measures trade flows and reflects an economy's international competitiveness. If a country is running a current account deficit, it means that its currency outflows are greater than its currency inflows, which generally means it is importing more than it is exporting. If one country is running a current account deficit, another country must be running a surplus, meaning that its currency inflows are greater than its outflows and it is generally exporting more than it is importing.

Look carefully at the current account balances from 2020 in Table 14.18. Over the last two decades, countries such as China and Germany have run persistent current account surpluses. They have successfully exported more products than they have imported, and they have experienced strong export-led growth. In contrast, the UK and USA have run persistent current account deficits, largely driven by very large deficits on the balance of trade in goods. The current account deficits have been mainly financed by borrowing.

Table 14.18 Current account balance of selected countries, 2020 (£m)

China	+273,980
Germany	+269,077
United Kingdom	–55,888
United States	–616,087

Sources: World Bank data, ONS

Do current account deficits pose problems?

While a short-run deficit or surplus on current account does not pose a problem, a persistent or long-run imbalance indicates fundamental disequilibrium. However, the nature of any resulting problem depends on the size and cause of the deficit: the larger the deficit, the greater the problem is likely to be. The problem is also likely to be serious if the deficit is caused by the uncompetitiveness of the country's industries. Although in the short run a deficit allows a country's residents to enjoy living standards boosted by imports, and thus higher than would be possible from the consumption of the country's output alone, in the long run, the decline of the country's industries in the face of international competition lowers living standards.

In a poor country, a current account deficit can be justified because of the country's need to import capital goods on a large scale to modernise the country's infrastructure and to promote economic development.

Do current account surpluses pose problems?

While many people agree that a persistent current account deficit can pose serious problems, few realise that a balance of payments surplus on current account can also lead to problems. Because a surplus is often seen as a sign of national economic success, a popular view is that the bigger the surplus, the better must be the country's performance.

Insofar as the surplus measures the competitiveness of the country's exporting industries, this is obviously true. There are, nevertheless, reasons why a large payments surplus is undesirable, though a small surplus may be a justifiable objective of government policy.

One country's surplus is another country's deficit

The balance of payments must balance for the world as a whole, so it is impossible for all countries to run surpluses simultaneously. Unless countries with persistently large surpluses agree to take action to reduce their surpluses, deficit countries cannot reduce their deficits. Deficit countries may then be forced to impose import controls from which all countries, including surplus countries, eventually suffer. In an extreme scenario, a world recession could be triggered by the resulting collapse of world trade.

At various times since the 1970s, the current account surpluses of the oil-producing countries have led to this problem, as have Japan's and more recently China's payments surpluses, which have been the counterpart to the US trade deficit. On several occasions, the US government has faced pressure from US manufacturing and labour interests to introduce import controls and other forms of protectionism. When introduced, US protectionism undoubtedly harms world trade.

Non-oil-exporting developing countries, almost without exception, also suffer chronic deficits, although these are very different from the US trade deficit. The imbalance of trade between more developed and less developed economies cannot be reduced without the industrialised countries of the Global North taking action to reduce surpluses which have been gained at the expense of the developing economies of the Global South.

A balance of payments surplus can be inflationary

A balance of payments surplus can be an important cause of domestic inflation because it is an injection of aggregate demand into the circular flow of income, which increases the equilibrium level of nominal or money national income. If there are substantial unemployed resources in the economy, this has the beneficial effect of reflating real output and jobs. However, if the economy is initially close to full capacity, demand-pull inflation results.

Three factors that influence a country's current account balance

We have made passing reference throughout section 14.3 to factors which might influence a country's current account balance. Three of these are productivity, inflation and the exchange rate, though primary income flows also need to be considered.

- **Productivity.** Improving labour productivity, or output per worker, is critical to the success of supply-side policies intended to improve both the price competitiveness and the quality competitiveness of a country's exports in international markets (see the later sections 'Supply-side policies and the current account' and 'Export-led growth and the UK economy' on pages 562–63).
- **Inflation.** The key point here is not a country's *absolute* rate of inflation, but its rate of inflation *relative* to those of its trading competitors. If a country's inflation rate is higher than the inflation rates of competitor nations, the country's exports will lose their price competitiveness, which will lead to a deterioration in the country's current account balance. However, the price elasticities of demand for exports and imports may complicate things, as will be explained below.
- **Exchange rate.** Changes in the exchange rate of a country's currency have a rather similar effect to changes in its relative rate of inflation. In short, a rising exchange rate increases the foreign currency prices of the country's exports and reduces their competitiveness. Meanwhile, imports become more price competitive. The price elasticities of demand for exports and imports are relevant to the analysis of the effects of a change in a country's exchange rate. Note also that, with respect to a fall in the exchange rate, if a rise in the domestic inflation rate relative to that of other countries exactly matches the fall in its exchange rate, the two effects cancel each other out, and the country's international price competitiveness remains unchanged.

Policies to cure or reduce a balance of payments deficit

Traditionally, it was thought that a government (or its central bank) can use three different policies to try to cure a persistent deficit caused by an over-valued exchange rate. These are the '3 Ds' of deflation, direct controls, and devaluation or currency depreciation, which are shown in Figure 14.16.

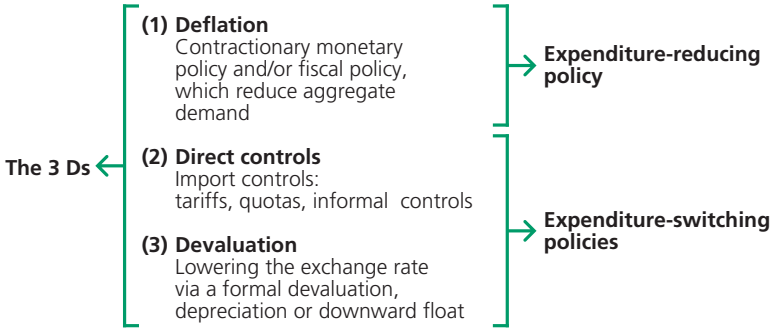


Figure 14.16 The '3 Ds' of deflation, direct controls and devaluation

A deflationary policy, which in this context refers to a reduction in the level of aggregate demand in the economy, reduces a current account deficit because for the most part it is an example of an **expenditure-reducing policy**. As aggregate demand falls, people's incomes fall and hence spending on imports is reduced. By contrast, import controls and devaluation are for the most part examples of **expenditure-switching policies** and affect the balance of payments by changing the relative price of home-produced and foreign goods.

KEY TERMS

expenditure-reducing policy a government policy which aims to eliminate a current account deficit by reducing the demand for imports, by reducing the level of aggregate demand in the economy. Conversely, to reduce a current account surplus, aggregate demand would be increased and spending on imports would rise.

expenditure-switching policy a government policy which aims to reduce a current account deficit by switching domestic demand away from imports to domestically produced goods. Conversely, to reduce a current account surplus, the policy would aim to switch domestic demand away from domestically produced goods towards imports.

STUDY TIP

Besides the '3 Ds' of deflation, direct controls and devaluation, you should also consider how supply-side policies may be needed to improve the long-term competitiveness of UK industries. Knowing how to correct (reduce or eradicate) a current account deficit is a very commonly assessed exam question.

Deflation as an expenditure-reducing policy

Deflationary policy involves using contractionary monetary and/or fiscal policy to reduce the demand for imports. For example, if the marginal propensity to import in the economy is 0.4, reducing aggregate demand by £10 billion should cause spending on imports to fall by £4 billion. This is an expenditure-reducing policy.

Although deflation is primarily an expenditure-reducing policy, it also has an expenditure-switching element. By reducing the rate of domestic price inflation relative to inflation rates in other countries, deflation improves the price competitiveness of exports and reduces that of imports.

However, in modern economies this is usually quite a small effect, and the main effect of deflationary policies, at least in the short run, is to reduce aggregate demand and depress economic activity in the domestic economy.

Output, incomes and employment tend to fall rather than the price level. Falling demand for domestic output may force firms to seek export orders, so as to use spare production capacity. However, because exports are generally less profitable than domestic sales, a sound and expanding home market may be necessary for a successful export drive.

In summary, when deflating aggregate demand to achieve the external objectives of supporting the exchange rate and reducing a current account deficit, a government sacrifices the domestic economic objectives of full employment and economic growth. For this reason, governments may choose

to use expenditure-switching policies of import controls and devaluation, in preference to expenditure-reducing deflationary policies. However, when an economy is overheating, expenditure-reducing measures are more appropriate than expenditure-switching policies. Reducing aggregate demand will help to control domestic inflationary pressures as well as to correct a balance of payments deficit.

SYNOPTIC LINK

In Chapter 10 we examined the components of aggregate demand and took a first look at policies such as fiscal policy that can be used by government to deflate the economy. Remember that deflation is a policy used to reduce aggregate demand — though it more often refers to a fall in the price level.

Direct controls as expenditure-switching policies

Direct controls involve imposing quotas or even outright bans (embargoes) on imports. These directly cut or prevent expenditure on imports and, as a result, people switch their spending from foreign to home-produced goods. Together with import duties or tariffs, which make imports less price competitive, direct controls do not, however, cure the underlying cause of a current account deficit, namely the uncompetitiveness of a country's goods and services. Moreover, because a country essentially gains an advantage at the expense of other countries, import controls tend to provoke retaliation.

Free-market economists believe that protectionism reduces specialisation and causes world trade, world output and economic welfare to fall. Because of this, international organisations such as the European Union and the World Trade Organization have reduced the freedom of individual countries to impose import controls unilaterally to improve their current accounts. However, the EU does use its common external tariff to provide protection for *all* its member states. As a member of the WTO, the UK government is unlikely to be able to use import controls to reduce a current account deficit.

However, the rules-based system of international trade has come under strain since 2016. The election of Donald Trump as US president resulted in new trade wars with China and the threatened withdrawal of the USA from the WTO. In 2022, Russia's invasion of Ukraine led to the international community imposing economic sanctions on trade with Russia — though not all countries were willing participants in these sanctions. The UK's withdrawal from the EU also led to difficulties in establishing the trading rules between Northern Ireland and the EU.

Devaluation or currency depreciation as an expenditure-switching policy

The exchange rate is the value of one currency in terms of another currency. The word *devaluation* is used in a number of different ways. In a narrow sense, a country devalues by reducing the value of a fixed exchange rate. (Fixed exchange rates are explained in section 15.1.) However, the term is sometimes used in a looser way to describe any fall in the value of the currency — that is, a decrease in the exchange rate. If the exchange rate is floating, then a fall in its value is usually referred to as a depreciation of the exchange rate (though devaluation can be used here as well).

The word 'depreciation' can also confuse. Devaluation or a downward float causes an *external* depreciation of the currency; more units of the currency are needed to buy a unit of *another* currency. This should not be confused with an *internal* depreciation of the currency, occurring when there is inflation *within* the economy and the internal purchasing power of the currency falls as the price level rises.

In a fixed exchange rate system, an increase in the external value of a currency is known as a *revaluation*, whereas in a floating rate system it is known as an *appreciation* in the exchange rate.

If we assume the unavailability of import controls, if a country wishes to reduce a current account deficit, it must generally choose between *deflation* and *devaluation*, or *currency depreciation*. As with tariffs and export subsidies, a fall in the external value of the currency has a mainly expenditure-switching effect. By increasing the price of imports relative to the price of exports, a successful devaluation switches domestic demand away from imports and towards home-produced goods. Similarly, overseas demand for the country's exports increases in response to the fall in export prices.

Price elasticity of demand and devaluation

The effectiveness of a fall in the exchange rate in reducing a balance of payments deficit depends to a significant extent on the price elasticities of demand for exports and imports.

As Figure 14.17 shows for the UK, when the demand for exports and the demand for imports are both highly price elastic, a fall in the exchange rate *is likely to* reduce a current account deficit.

Following a devaluation, the domestic price of imports into the UK rises from P_1 to P_2, while the overseas price of UK's exports falls from P_3 to P_4. As a result, UK residents spend less on imported goods following an increase in their relative prices. At the same time, residents of overseas countries spend more on the UK's exports, whose relative prices have fallen.

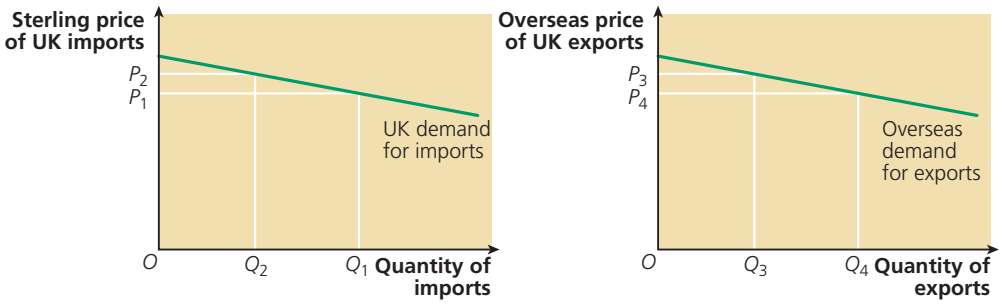


Figure 14.17 The effect of a devaluation (or downward float) of the pound's exchange rate on the current account of the UK balance of payments

On the import side, in Figure 14.17, spending falls from the rectangular area bounded by P_1 and Q_1 to the area bounded by P_2 and Q_2. Higher import prices mean that consumers switch to the now cheaper domestically produced substitutes.

In a similar way, expenditure on the country's exports increases from the rectangular area bounded by P_3 and Q_3 to the area bounded by P_4 and Q_4. Overall, the current account improves, assuming the demand for imports and the demand for exports are both price elastic.

SYNOPTIC LINK

Make sure you can apply the microeconomic concept of elasticity to macroeconomic analysis of the expenditure-switching policy of devaluation. Knowing the price elasticities for imports and exports will tell you whether a devaluation will 'work' in reducing the current account deficit.

EXTENSION MATERIAL

The Marshall–Lerner condition

It is more difficult to see what may happen to the current account when, for example, the demand for exports is price inelastic but the demand for imports is price elastic. Fortunately, the *Marshall–Lerner condition* provides a simple rule to assess whether a change in the exchange rate can improve the current account. The condition states that when the *sum* of the export and import price elasticities is greater than unity (unity refers to 1) (ignoring the minus sign), a fall in the exchange rate can reduce a deficit and a rise in the exchange rate can reduce a surplus. When, however, the export and import price elasticities of demand are both highly inelastic, summing to less than unity, a fall in the exchange rate can have the perverse effect of worsening a deficit (while a revaluation might increase a surplus).

The Marshall–Lerner condition is a *necessary*, but not *sufficient* condition for a fall in the exchange rate to reduce a payments deficit. For a devaluation or currency depreciation to be successful, firms in the domestic economy must have spare capacity with which they can meet the surge in demand brought about by the fall in the exchange rate. This means that if the economy is working at, or close to, full capacity, expenditure-reducing deflation and expenditure-switching devaluation should best be regarded as complementary policies rather than as substitute policies for reducing a current account deficit. Deflation *alone* may be unnecessarily costly in terms of lost domestic employment and output, yet may be necessary to provide the spare capacity and conditions in which a falling exchange rate can successfully cure a payments deficit. It may also be needed to offset the potential inflationary consequences of a fall in the exchange rate, which will erode the improvement in competitiveness resulting from the initial fall in the external value of the currency.

EXTENSION MATERIAL

How changes in exchange rates affect the prices of imports

To understand why, when the demand for exports and the demand for imports are both highly price elastic, a fall in the exchange rate *is likely to* reduce a current account deficit, consider the effect of a change in the exchange rate upon the UK prices of imports.

The UK firm, Quidsin supermarkets, buys 500 tonnes of Californian apples each month. The dollar price of the apples is $100 per tonne. The total amount that Quidsin has to pay for the apples, therefore, is $50,000. If the exchange rate in January is £1 = $1.50, then Quidsin will have to give up around £33,333 to acquire $50,000 (50,000 ÷ 1.50).

When the exchange rate changes a year later to £1 = $2.00, Quidsin now has to give up £25,000 to get the $50,000 that the supermarket chain needs to pay its suppliers. The strengthening of the pound's exchange rate has benefited Quidsin. The supermarket chain now has to give up fewer pounds to get the same amount of dollars. Quidsin's costs have fallen by around £8,333 a month.

However, two years later the situation changes again, with the exchange rate falling back to £1 = $1.50. This means that, once again, Quidsin has to give up £33,333 to acquire $50,000 (50,000 ÷ 1.50). Compared to a year earlier, Quidsin's costs have now risen by approximately £8,333.

The position for a UK business which *exports* rather than *imports* goods or services would be the exact opposite of the above.

EXTENSION MATERIAL

The J-curve effect

Even if domestic demand for imports and overseas demand for exports are both price elastic and spare capacity exists in the economy, in the short run the balance of payments may deteriorate. It takes time for demand to respond to the price changes, and firms within the country may still be unable immediately to increase supply following a fall in the exchange rate. In the short run, the Marshall–Lerner condition (explained in previous extension material) may not hold because elasticities of demand are lower in the short run than in the long run. In these circumstances, the balance of payments may worsen before it improves. This is known as the *J-curve effect*, which is illustrated in Figure 14.18.

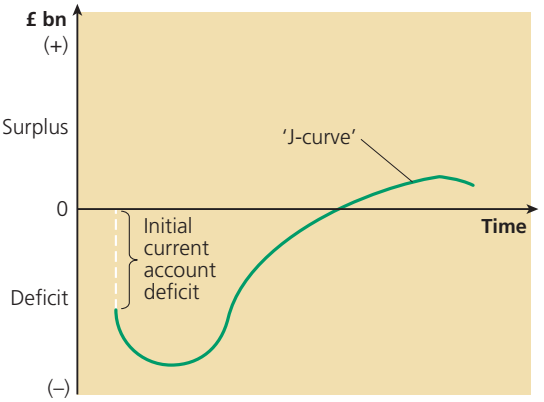


Figure 14.18 The J-curve effect

The initial worsening of the balance of payments that follows the fall in the exchange rate may reduce confidence in the idea that changing the exchange rate is the most appropriate method for reducing a deficit in the current account of the balance of payments. Falling confidence may, in turn, cause capital outflows to occur that destabilise both the balance of payments and the exchange rate. The J-curve effect thus reduces the attractiveness of exchange rate adjustment as an instrument to correct payments disequilibrium. Even when the benefits of a falling exchange rate are realised, they may be short-lived. The increased price competitiveness produced by the devaluation is likely to be eroded as increased import prices raise the country's inflation rate.

Nevertheless, if conditions are right, devaluation can reduce a current account deficit. Despite occurring on 'Black Wednesday', the pound's devaluation in September 1992 was extremely successful, at least for a number of years. There were two main reasons for this. First, expenditure reduction in the severe recession of the early 1990s created the spare capacity that enabled successful expenditure switching following the pound's devaluation. Secondly, the adoption and achievement of a low inflation target helped to prevent the benefits of the fall in the exchange rate being eroded by accelerating inflation. It was also believed that the recession of the early 1990s was so severe that it lowered expectations of future inflation, which also helped export price competitiveness.

In addition, the factories built in the UK by Japanese companies such as Honda and Toyota had just come on stream, producing goods of a quality that people wanted, in the UK and overseas. Some economists believe that the supply-side reforms that were introduced in the 1980s and 1990s also helped.

In contrast, the pound's devaluation of 2008 did not improve the UK's current account position. Despite the pound losing over 25% of its value against its major trading partners, the current account deficit did not shrink. Instead by 2010 the UK's deficit on the current account was larger than it had been in 2007. This was partly because in the years following the depreciation, the UK's major trade partners were in a deep recession, making it difficult for exporters to sell in these markets. Hence, a devaluation of a country's currency is in itself insufficient to eliminate a current account deficit because other macroeconomic factors are important.

Supply-side policies and the current account

Deflation, devaluation and direct controls (the '3 Ds') may be effective *short-term* policies for reducing current account deficits. However, it is now increasingly recognised that *long-term* improvement in trade flows requires appropriate and successful supply-side policies and supply-side improvements undertaken by firms within the economy.

The ability of the UK economy to deliver sustained growth of exports and meet the challenge of imported goods and services depends on making UK exports *quality competitive* as well as *price competitive*. A low exchange rate, low interest

rates at which UK firms can borrow and low domestic inflation all contribute to increased export price competitiveness. However, price competitiveness on its own is not enough. UK goods and services must also be quality competitive. This involves good design and well-made products.

Improved quality competitiveness may only be achievable in the long run if helped by appropriate government supply-side policies together with supply-side improvements undertaken by the private sector. Supply-side policies which promote greater investment in research and development and improved marketing strategies can have powerful long-term effects in improving quality competitiveness. Supply-side improvements, followed by an outward shift of the *LRAS* curve, provide the economy with increased capacity — enabling a reallocation of resources towards exporting. Successful supply-side policies can also improve exports and lead to import substitution, by increasing labour productivity, which in turn is likely to improve the price competitiveness of UK exports.

SYNOPTIC LINK

Chapter 13 has already explained supply-side theory and supply-side policies in some depth. For this section of the course, think about how supply-side policies can improve the performance of export industries, or those UK industries producing goods that we often bypass in favour of purchasing an imported substitute.

Export-led growth and the UK economy

Export-led growth is a term first introduced about 45 years ago, in the context of speeding up the rate of economic growth in developing countries. The term is still widely used with regard to growth strategies in the developing world, but in recent years in the UK, together with the concept of investment-led growth, export-led growth has been more narrowly focused on how best to increase the competitiveness of UK exports and achieve sustained economic growth, without the UK economy slipping into recession. Perhaps not surprisingly, debate in the UK about how to achieve export-led growth became prominent when the economy appeared to be stuck in recession in 2008 and 2009, and there seemed to be little hope of speedy economic recovery.

To remind you, short-term economic growth occurs whenever one or more of the components of aggregate demand increases. The components of aggregate demand are consumption (C), investment (I), government spending (G) and net export demand ($X - M$). In recent decades, short-term economic growth has been largely consumption-led, though in 2008 and 2009, via its fiscal stimulus, the Labour government tried to achieve government spending-led growth to 'spend the UK economy out of recession' as part of the response to the collapse in large parts of the financial sector in the UK.

Consumption-led growth

Consumption-led growth leads to two big problems. First, since much consumer spending is on goods and services produced in other countries, consumption-led growth 'sucks' imports into the economy, which may soon lead to a worsening deficit in the current account of the balance of payments. A growing payments deficit may then force the government and the Bank of England to deflate aggregate demand, which at best slows the rate of economic growth, and at worst triggers recession.

Secondly, and perhaps more significant, recent economic history has shown that consumption-led growth is unsustainable. This is because a rapid growth of consumption, fuelled by increased household debt and consumer borrowing, leads to speculative bubbles, particularly in the housing market. When these bubbles are eventually 'pricked', aggregate demand collapses, bringing to an end the 'boom' phase of the economic cycle. The rate of economic growth declines and perhaps becomes negative.

In this scenario, export-led growth and investment-led growth are seen as 'magic bullets' (factors that create positive effects) which, if they can be achieved, will increase productivity, competitiveness and supply-side reform within the economy, and allow sustained economic growth to take place. In such a situation, short-term economic growth, brought about by the increase in aggregate demand, will seamlessly move into sustainable long-term growth. This is because export-led growth and investment-led growth are associated with increased productivity and with the modernisation and enlargement of the economy's productive capacity.

Achieving export-led growth and investment-led growth is, however, much easier said than done. At the time of writing this chapter, the UK's continuing recovery from recession appears once again to be consumption-led, with little or no evidence of export-led or investment-led growth. Whether this heralds the eventual onset of another recession remains to be seen, and it must be remembered that adverse economic 'shocks' in the wider world economy can also lead to a collapse in aggregate demand in the UK economy.

Government policies to correct or reduce a balance of payments surplus on current account

The policies available to a government for reducing a balance of payments surplus are the opposite of the '3 Ds' of deflation, direct controls and devaluation, which are appropriate for correcting a payments deficit. The policies are the '3 Rs' of reflation, removal of import controls and revaluation:

- **Reflation.** Reflating demand, via expansionary monetary policy or fiscal policy, increases a country's demand for imports.
- **Removing import controls.** Trade can be liberalised by removing import controls.
- **Revaluation.** There have been calls on countries with large payments surpluses, such as Japan and China, to revalue in order to reduce global payments imbalances. But because there is much less pressure on a surplus country to revalue than on a deficit country to devalue, such calls have not usually been successful. It is also worth noting that, for a revaluation to reduce a current account surplus, the Marshall–Lerner condition must be met. In addition, a reverse J-curve, illustrated in Figure 14.19, may operate, causing the payments surplus to get bigger immediately after the revaluation, before it eventually starts to get smaller.

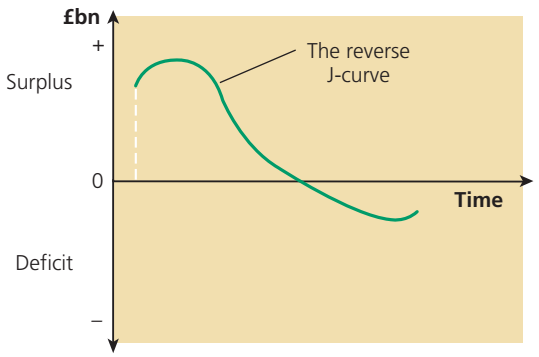


Figure 14.19 The reverse J-curve effect

The significance of balance of payments deficits and surpluses for individuals

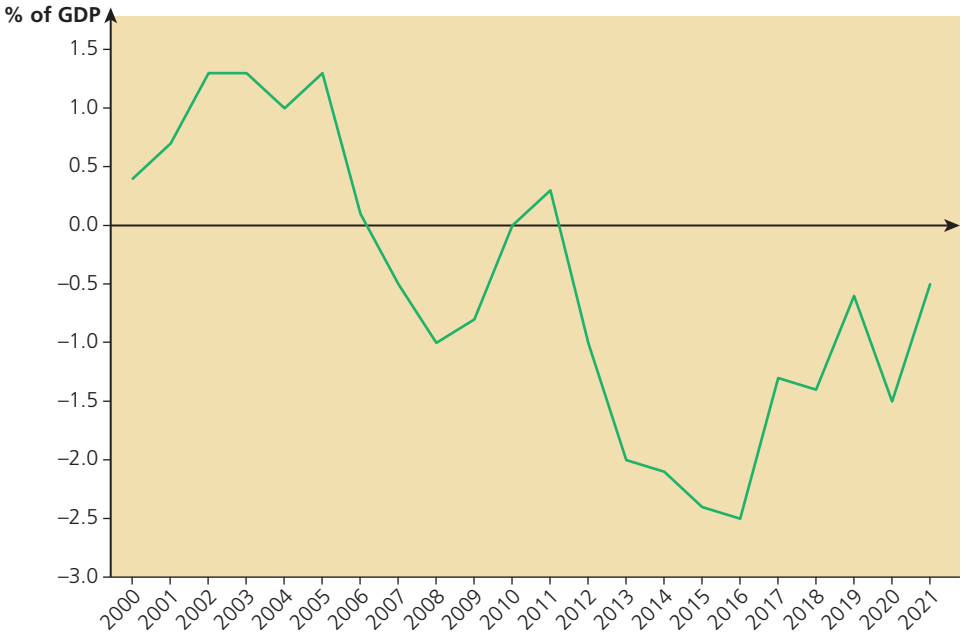
Earlier in this section we addressed the issue of whether balance of payments deficits and surpluses on current account pose problems for the country experiencing the deficit or surplus.

It was explained that, while a short-run deficit or surplus on current account does not pose a problem, a persistent or long-run imbalance indicates fundamental disequilibrium. In the case of a deficit, the nature of any resulting problem depends on the size and cause of the deficit: the larger the deficit, the greater the problem is likely to be. The problem is also likely to be serious if the deficit is caused by the uncompetitiveness of the country's industries. Although in the short run a deficit allows a country's residents to enjoy living standards boosted by imports, and thus higher than would be possible from the consumption of the country's output alone, in the long run, the decline of the country's industries in the face of international competition lowers living standards.

While many people agree that a persistent current account deficit can pose serious problems, we also said that few realise that a balance of payments surplus on current account can also lead to problems. As a surplus is often seen as a sign of national economic success, a popular view is that the bigger the surplus, the better must be the country's performance. This is obviously true insofar as the surplus measures the competitiveness of the country's exporting industries. However, although a small surplus may be a justifiable objective of government policy, a large payments surplus should be regarded as undesirable for two reasons:

- One country's surplus is another country's deficit. Since we do not trade with Mars or any other planet, the balance of payments must balance for the world as a whole. It is therefore impossible for all countries to run surpluses simultaneously. Unless countries with persistently large surpluses agree to take action to reduce their surpluses, deficit countries cannot reduce their deficits. Deficit countries may then be forced to impose import controls from which all countries, including surplus countries, eventually suffer.
- A balance of payments surplus is an injection of aggregate demand into the circular flow of income and this can be inflationary. If there are substantial unemployed resources in the economy, this can have the beneficial effect of reflating real output and jobs. However, if the economy is initially close to full capacity, demand-pull inflation results.

We noted earlier that, in recent years, the UK's current account deficit has increased in size. This has not necessarily been regarded as a problem because it has been financed by inward capital flows. However, the deficit has been increasing as a percentage of national income and some economists believe that, unless the deficit is reduced, it could lead to capital outflows as foreign investors question the ability of the UK to finance its growing overseas debts. It is not a deterioration in the balance of trade in goods and services that has been the main problem in recent years. As Figure 14.20 shows, a significant cause of the growing current account deficit is a deterioration in the primary income balance.



Source: ONS

Figure 14.20 The changes in the UK's primary income balance, 2000–21

Primary income is the balance between the income (i.e. profits, dividends and interest payments) received on the UK's foreign investments, and the income paid to overseas investors on their UK investments. Figure 14.20 shows that since 2011 there has been a significant deterioration of the UK's primary income position, which has widened the deficit on the current account. According to **www.parliament.uk**:

> ... the deterioration in the primary account was driven by UK residents receiving lower income on their overseas foreign direct investments (FDI). This in turn was a result of both lower returns on those investments (thanks in part to economic weakness in the eurozone), and a reduction in the total stock of UK overseas FDI. In contrast, foreign residents have continued to expand their holdings of UK assets (something that may be linked to the relative strength of the UK's recovery and its status as a 'safe haven'), and have shifted the composition of their assets from low-yielding debt to riskier, higher-yielding equity, thereby increasing the rate of return on their holdings.

Implications for the global economy of other economies taking action to reduce their current account imbalances

As we have seen, countries are more likely to take action to try to reduce a current account deficit than to reduce a current account surplus. Indeed, surplus countries very often like to keep things exactly as they are, particularly if they benefit from artificial price competitiveness caused by an under-valued exchange rate.

Perhaps not surprisingly, corrective action by an economy to reduce the size of a current account deficit may have very little effect on other economies, first, if the deficit is small, and secondly, if the economy forms only a tiny part of the world economy. The same is not true, however, if the economy taking the corrective action is large.

The USA

The USA is the country which runs the largest current account deficit, importing far more than it exports. Many of these imports are cheap Chinese manufactured goods and components that US firms such as Apple design before they are assembled into finished consumer goods (often in China). China likes it this way because the USA provides a ready market for its exports. Many US importing firms and consumers also like it this way because they benefit from cheap components and consumer goods. (However, other US manufacturing firms and the workers they employ do not like it this way. Unless these firms move out of the USA and relocate in countries with cheap labour, they may be forced to go under, and either way, American workers will lose their jobs.)

To correct the deficit and to save American jobs, the US government may resort to protectionism. In the past the USA used its power to engage in various forms of covert or hidden protectionism, such as giving preferential defence contracts to aircraft manufactures and subsidies to farmers. Such policies undoubtedly harm countries which sell goods to the USA. If a smaller and less powerful country were to take this action, it might end up suffering from retaliation from other economies, but the size and power of the USA generally make this impractical. In recent years the USA has aggressively demanded better trading terms by breaking WTO rules and introducing tariffs on imported goods from China and the EU. It has upset traditional allies Mexico and Canada and demanded the renegotiation of NAFTA to better suit the interests of the USA.



The USA's large current account deficit results from high imports of manufactured goods, such as smartphones assembled in China

The USA has never really deflated its economy in order to correct its payment deficit, but when the US economy does suffer from a downturn — for example, recession in 2008 and 2009 — other countries have faced a sudden loss of export markets. There is a saying: 'When America sneezes, the rest of the world catches a cold.' The size of the US economy means that countries which export to the USA are highly dependent on it continuing to grow. A recession in the USA tends to mean that other countries import recession from it, which is clearly not what those countries want.

Competitive devaluation

As we have seen, another action a country can take to reduce a current account deficit is to devalue or encourage a depreciation of its exchange rate. Again, if this policy is viewed by other countries as an attempt to gain at their expense, it might lead to retaliatory devaluations and to an exchange rate war from which nobody benefits. Encouraging the dollar to fall has not always been an effective way of reducing the USA's balance of payments deficit. This is because other currencies, such as China's renminbi (RMB) or yuan, have at times been effectively 'pegged' to the dollar, which means that if the dollar's exchange rate falls, so does the RMB's exchange rate. (Note that the official name for China's currency is the renminbi, but when talking about quantities of the currency it is referred to in terms of the number of yuan; this is similar to 'sterling' being the official name for the British currency but 'pounds' being used to measure the quantity of the British currency.) However, China stunned financial markets in August 2015 by devaluing the RMB on two consecutive days in order to counter a slowdown in China's growth rate. This is despite the fact that China had previously been reluctant to raise the RMB's exchange rate and lose its price competitive advantage. This was repeated in 2019, leading to accusations by President Trump of the Chinese government being a 'currency manipulator'.

In the case of the USA and China, the imbalance in the two countries' current accounts is largely resolved by large capital flows from China into the USA. For example, China has purchased large quantities of US government bonds. Today, China holds US Treasury bonds worth more than $1 trillion. These capital flows finance the US trade deficit. Indeed, the world as a whole is generally willing to 'invest' in the USA, irrespective of the state of the US current account, because the US dollar is a reserve currency and is viewed as a 'safe haven'.

Smaller countries

For a smaller country such as the UK, a downward float of the pound's exchange rate might be a viable method of reducing a current account deficit. A fall in the pound's exchange rate would probably have relatively little effect on other economies. However, the effectiveness of a depreciation would depend on how big it was, on whether it provoked retaliation, and on the price elasticities of demand for the UK's exports and imports. A fall in the exchange rate might also provoke a 'hot money' flow out of the pound, leading to a further fall in the value of the currency that might then destabilise the UK economy.

SECTION 14.3 SUMMARY

- The balance of payments is the part of the national accounts that measures all the currency flows into and out of the country in a particular time period.
- The two main parts of the balance of payments are the current account and the financial account, which records most of the capital flows between countries.
- The financial account can be divided into direct investment, portfolio investment and short-term capital flows, which include many speculative or 'hot money' flows.
- There are two other, much smaller sections of the balance of payments: the 'capital account' and 'net errors and omissions'.
- 'Hot money' flows can destabilise the exchange rate, the balance of payments and indeed the whole economy.
- The expenditure-reducing policy of deflation and/or the expenditure-switching policies of import controls and devaluation can be used to reduce a current account deficit.
- Sometimes deflation and devaluation should be used in tandem.
- Likewise reflation, removal of import controls and revaluation can be used to reduce a surplus.
- Supply-side policies can also help to reduce a current account deficit.

Questions

1 Do you agree that, for a developed economy such as the UK, the advantages of globalisation exceed any disadvantages? Justify your answer.

2 Explain the difference between absolute advantage and comparative advantage.

3 Assess the view that the progress made towards free trade has brought significant economic benefits both to the global economy and to the UK.

4 Explain the policies that may be used to reduce a deficit on the current account of the balance of payments.

5 Do you agree that supply-side reforms are needed if the UK's current account is to improve in the long run? Justify your answer.

6 Using a demand and supply diagram, explain how a trade deficit is eliminated in a freely floating exchange rate system.



15 Exchange rate systems and economic growth and development

Following on from Chapter 14, this is the second chapter in the book on international economics. Section 15.1 explains exchange rate systems. Exchange rate equilibrium cannot be fully understood without understanding the balance of payment equilibrium, and vice versa. The former is measured in terms of the relative prices of currencies, whereas the latter is stated in terms of the quantities of goods, services and financial flows between countries. Chapter 14 explained the balance of payments, but this chapter focuses on exchange rates. Section 15.2, the final part of the book, is significantly different, surveying the topic of development economics and some of the problems faced by the world's poorer economies.

LEARNING OBJECTIVES

These are to understand:

- how exchange rates are determined in a freely floating exchange rate system
- how a government and central bank can intervene to manage an exchange rate
- the advantages and disadvantages of freely floating and fixed exchange rate systems
- advantages and disadvantages of joining a currency union
- the difference between economic growth and economic development
- the main characteristics of less developed economies
- the indicators of economic development
- the factors that affect growth and development
- the barriers to growth and development
- the policies that might be adopted to promote economic growth and development
- the role of aid and trade in promoting growth and development.

15.1 Exchange rate systems

Introduction to exchange rates

So far in this book, we have made several mentions of exchange rates, particularly in the context of the balance of payments. This should not be surprising: whenever we discuss exchange rates, we usually have to make reference to the balance of payments, and vice versa.

Although domestic currencies are used to pay for internal trade within countries, imports are usually paid for in the currency of the country exporting the goods or services. An exchange rate measures how much of another currency a particular currency can buy; it is the external price of the currency quoted in terms of another currency. Exchange rates can also be measured against gold, or against a weighted average of a selection or 'basket' of currencies. Currencies are bought and sold in the foreign exchange market, which is now an international market, dominated by electronic trading. On a global scale, the market never closes, and ICT-based buying and selling takes place throughout the day and night.

It is normal in the UK to express the exchange rate as how much foreign currency one pound will buy. For example, if the exchange rate is £1 = €1.18, this shows how many euros we would get for each pound exchanged. A rise in the exchange rate means the pound buys a larger quantity of a foreign currency — this is also referred to as a currency strengthening, becoming stronger or appreciating in value. A fall in the exchange rate is where the pound buys a smaller quantity of a foreign currency (also known as a weakening, or a depreciation/devaluation).

QUANTITATIVE SKILLS 15.1

Worked example: using exchange rates

If the exchange rate is £1 = $1.50, what would $1 be worth in terms of £s?

To convert the exchange rate in this manner requires us to find the reciprocal of the rate.

In this case 1/1.50 = 0.6666, so $1 = £0.67 (to the nearest penny).

SYNOPTIC LINK

Chapter 11 mentioned the relationship between the exchange rate and a country's international competitiveness, and discussed how a falling exchange rate increases import prices and causes cost-push inflation.

Chapter 12 explained the link between the exchange rate and interest rates, in the context of monetary policy.

Chapter 14 looked at globalisation, multinational corporations, trade and the balance of payments. Changes in the value of an exchange rate can have a major effect on all of these topics. The appreciation or depreciation of a currency affects economic activity because it changes the price of exports and imports.

The meaning and measurement of an exchange rate

KEY TERM

exchange rate the external price of a currency, usually measured against another currency.

These days the **exchange rate** of a currency is simply the external price of the currency in terms of another currency, such as the US dollar. The convention of quoting exchange rates in terms of the US dollar dates from the years after the Second World War. Before 1914 most exchange rates were expressed in terms of gold and only after 1945 did the dollar become the near universally accepted standard by which the external values of other currencies were measured.

In recent years, in response to the changing pattern of UK trade, the pound's exchange rate is as often quoted against the euro as it is against the dollar. The sterling exchange rate index (ERI) is also used to measure the pound's exchange rate. The ERI does not measure the pound's external value against a particular currency. Rather it is a trade-weighted average of the pound's exchange rate against a number of leading trading currencies, calculated to reflect the importance of each currency in international trade. Over time the UK may see a gradual reduction in trade with the European Union (EU) now it is no longer a member, and a rise in trade with other countries.

EXTENSION MATERIAL

The real exchange rate

The different exchange rates mentioned so far are all *nominal* exchange rates. These must not be confused with a *real* exchange rate, which measures the weighted average value of a country's currency relative to a basket of other major currencies, adjusted for the effects of inflation. The real exchange rate of the pound, which is a measure of competitiveness, is calculated by the following formula:

$$\text{pound's real exchange rate} = \text{sterling index} \times \frac{\text{index of the domestic price level}}{\text{index of weighted foreign price levels}}$$

For simplicity, assume that the domestic price level rises by 10%, the foreign price level remains unchanged and the nominal value of the domestic currency depreciates by 10%; in that case, the real exchange rate remains unchanged. The price competitiveness of the country's products has not changed, and the improvement in competitiveness brought about by the fall in the nominal exchange rate has been exactly offset by a higher rate of domestic inflation.

QUANTITATIVE SKILLS 15.2

Worked example: calculating percentage changes in an exchange rate from data presented in index number form

Table 15.1 shows the value of a country's currency against other currencies, in index number form.

Table 15.1 Exchange rate index, 2020–23

Year	Exchange rate index (2021 = 100)
2020	96
2021	100
2022	106
2023	102

Calculate the percentage change in the country's exchange rate between: (a) 2020 and 2023, (b) 2021 and 2022.

(a) Between 2020 and 2023 the index changed by 6 points. The percentage change between these years was therefore (6/96) × 100, which is a 6.25% increase.

(b) In this example, because the comparison is with the base year index number (100), the change in index points and the percentage change are the same: a 6-point index number change and a 6% change.

TEST YOURSELF 15.1

Explain the difference between a currency's *nominal* exchange rate and its *real* exchange rate.

The different types of exchange rate system

KEY TERM
managed exchange rates where the government intervenes in the foreign exchange market to influence its value.

Figure 15.1 shows the main types of exchange rate system. The two extreme types are freely floating exchange rates (also known as cleanly or purely floating exchange rates) and rigidly fixed exchange rates. A fixed exchange rate is the most extreme form of **managed exchange rates**.

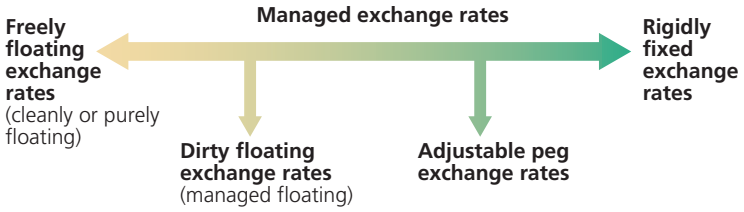


Figure 15.1 The different types of exchange rate system

The managed exchange rates, which lie between the extremes of freely floating and rigidly fixed exchange rates, take two main forms: adjustable peg and managed-floating (or dirty-floating) exchange rates. Adjustable peg exchange rates resemble fixed exchange rates in many respects, but the rate at which the exchange rate is fixed may be changed from time to time. A formal devaluation reduces the fixed exchange rate, while revaluation increases the fixed rate. With an adjustable peg system, there is usually a band in which the exchange rate is allowed to fluctuate.

Freely floating exchange rates

KEY TERM
freely floating exchange rate the exchange rate is determined solely by the interplay of demand for, and supply of, the currency.

In a regime of **freely floating exchange rates**, the external value of a country's currency is determined on foreign exchange markets by the forces of demand and supply alone. Towards the end of this section, we will remind you of how, in recent decades, capital flows and speculation have been extremely significant in influencing the supply of and demand for a currency, and hence its exchange rate.

However, we will first simplify by assuming that a currency is demanded on foreign exchanges solely for the payment of trade and that trade flows alone determine exchange rates. We will also assume that any holdings of foreign currencies surplus to the immediate requirement of paying for trade are immediately sold on the foreign exchange market.

Explaining the slope of the demand and supply curves for pounds

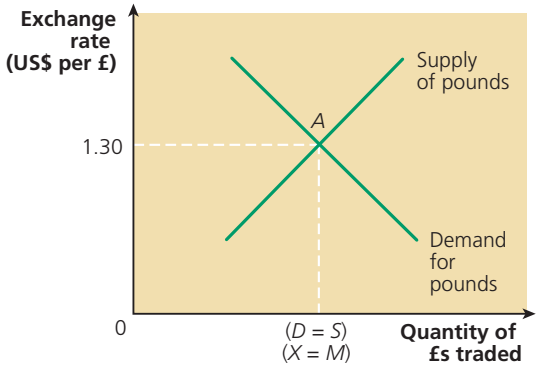


Figure 15.2 Exchange rate equilibrium in a freely floating exchange rate system

When the exchange rate of the pound falls, UK exports become more competitive in overseas markets. The volume of UK exports increases, leading to greater overseas demand for pounds to finance the purchase of these exports, provided that the demand for UK exports is price elastic. This explains the downward-sloping demand curve for pounds, which is illustrated in Figure 15.2.

But just as UK exports generate a demand for pounds on foreign exchange markets, so imports into the UK generate a supply of pounds. The explanation lies in the fact that UK trading companies generally pay for imports in foreign currencies. Importers must sell sterling on the foreign exchange market in order to purchase the foreign currencies needed to pay for the goods they are buying.

As the pound's exchange rate rises, fewer pounds are needed to buy a given quantity of foreign currency. This means that the sterling price of imports falls. UK consumers are likely to respond to the falling price of imports by increasing total spending on imports (which happens as long as the demand for imports is price elastic). A greater total quantity of sterling must be supplied on foreign exchange markets to pay for the imports — even though the sterling price of each unit of imports has fallen. The result is the upward-sloping supply curve of sterling depicted in Figure 15.2. This shows that, at higher exchange rates, more sterling is likely to be supplied on the foreign exchange market.

TEST YOURSELF 15.2

What is the pound's exchange rate against the US dollar and the euro at the time you are reading this chapter? Is it a floating or a fixed exchange rate, and has the exchange rate been rising or falling in recent months?

Exchange rate equilibrium in a freely floating exchange rate regime

Exchange rate equilibrium occurs at the market-clearing exchange rate at which the demand for pounds on foreign exchange markets equals the supply of pounds. In Figure 15.2, this is determined at point *A*. The equilibrium exchange rate is $1.30 to the pound.

At this exchange rate, the money value of exports (paid in sterling) equals the money value of imports (paid in foreign currencies). And provided we assume that exports and imports are the only items in the current account, the current account is also in balance.

We are assuming away any complications introduced by capital flows, so exchange rate equilibrium implies balance of payments equilibrium on current account and vice versa. As we stated in Chapter 14, the two equilibria are just different sides of the same coin: exchange rate equilibrium is price equilibrium, whereas current account equilibrium in the balance of payments (where $X = M$ in Figure 15.2) means that the quantity of the currency flowing into the country equals the quantity flowing out. Given the simplifying assumptions we have made, once the balance of payments is in equilibrium, there is no pressure for the exchange rate to rise or fall.

SYNOPTIC LINK

The value of a currency has a significant influence on the balance of the current account on the balance of payments. Refer back to Chapter 14, section 14.3, to help refresh your learning.

STUDY TIP

The value of the pound sterling is constantly fluctuating on the currency markets against the US dollar and the euro. Track the movements of the pound's exchange rates against these two currencies over the months since you started studying economics, and make sure you know when it appreciated or depreciated in value. Up-to-date knowledge enhances candidates' essay answers.

The adjustment process to a new equilibrium exchange rate

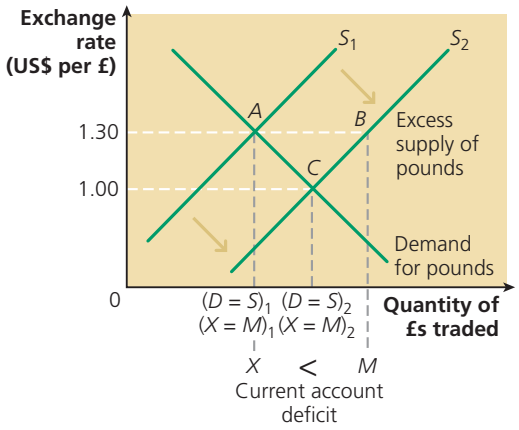


Figure 15.3 How a current account deficit is eliminated in a freely floating exchange rate system

We will now assume that some event or 'shock' disturbs the initial equilibrium — for example, an improvement in the quality of foreign-produced goods causes UK residents to increase demand for imports, whatever the exchange rate. In Figure 15.3, the increase in demand for foreign exchange to pay for imports causes the supply curve of the pound sterling to shift to the right from S_1 to S_2. (Remember, when more of foreign currencies is demanded, more pounds must be supplied on the foreign exchange market.) In the new situation, the current account of the balance of payments is in deficit by the amount ($X < M$) in the diagram — as long as the exchange rate stays at $1.30. At the $1.30 exchange rate, UK residents supply or sell more pounds than before to pay for imports, but because overseas residents still demand the same quantity of UK goods (assuming that their views on the quality of UK goods relative to foreign goods have not changed), the overseas demand for pounds to pay for UK exports remains at its previous level.

At the exchange rate of $1.30 to the pound, there is now an excess supply of pounds on the foreign exchange market, equal to the distance $B - A$. The market mechanism now swings into action to restore equilibrium — both for the exchange rate and for the balance of payments. When the excess holdings of pounds accumulated at the exchange rate of $1.30 are sold on the foreign exchange market, the pound's exchange rate falls. This increases the price competitiveness of UK exports while making imports less price competitive. The exchange rate falls until a new equilibrium exchange rate is reached at point C, where the exchange rate is $1.00 to the pound.

Note that the current account of the balance of payments is once again in equilibrium, but at $(X = M)_2$ rather than at $(X = M)_1$. This means that at the new equilibrium exchange rate ($1 to the pound), the money values of exports and imports are again equal in size, but the actual physical quantities of both exports and imports have increased.

Conversely, if the initial equilibrium is disturbed by an improvement in the quality of UK goods or services, the demand curve for the pound shifts to the right. This moves the current account into surplus, causing the pound's exchange rate to rise or appreciate in order to relieve the excess demand for pounds. Provided UK residents do not change their views on the relative quality of imports, the exchange rate rises until the balance of payments and exchange rate equilibrium are once again restored.

TEST YOURSELF 15.3

Aggregate demand in an economy increases at a time when its currency's exchange rate is initially in equilibrium. At this exchange rate, will the increase in aggregate demand create excess supply of the currency in the foreign exchange market? What is likely to happen to the exchange rate?

SYNOPTIC LINK

An exchange rate diagram explains how currency markets operate in the same way as the product market supply and demand diagrams in Chapter 3. It is advisable that you read over this material so that you don't confuse movements along a curve with changes in the conditions of demand and supply that cause a shift of a curve.

STUDY TIP

You should practise drawing demand and supply graphs to show a floating exchange rate. Remember to label the axes to reflect that the diagram is showing a currency market.

Advantages of floating exchange rates

Economists generally agree that, provided there are no distorting capital flows, freely floating exchange rates have the following advantages.

Automatically achieving balance of payments equilibrium

The exchange rate (which is the external price of the currency) should move up or down automatically to correct a payments imbalance. Provided the adjustment mechanism operates smoothly, a currency should never be over-valued or under-valued for long. Over-valued and under-valued exchange rates refer to the current exchange rate compared with what is believed to be its long-term purchasing power parity (PPP) exchange rate level (see Chapter 10). An over-valued currency is one where the market value of the exchange rate is currently above its PPP value. The opposite is the case for an under-valued exchange rate. In the event of an over-valued exchange rate causing export uncompetitiveness and a payments deficit, market forces should quickly adjust the exchange rate downwards towards its equilibrium price, which also achieves equilibrium in the balance of payments. Similarly, under-valuation should be quickly corrected by an upward movement of the exchange rate.

However, this adjustment process is based on the assumption that trade in exports and imports is the main determinant of the exchange rate's value. In today's world, many of the foreign currency transactions taking place are not for the purposes of trade but are for short-term speculative purposes (i.e. hot money). There are many instances where a country has both a rising deficit on the balance of trade but also a strong or even rising currency at the same time.

Improving resource allocation

If the world's resources are to be efficiently allocated between competing uses, exchange rates must be correctly valued. For efficient resource allocation in a constantly changing world, market prices must accurately reflect shifts in demand and changes in competitive and comparative advantage that result from technical progress and events such as discoveries of new mineral resources. In principle, a freely floating exchange rate should respond and adjust to these changes. By contrast, a fixed exchange rate may gradually become over-valued or under-valued, as demand or competitive and comparative advantage move against or in favour of a country's industries. Similarly, different rates of inflation between countries mean that, over time, fixed exchange rates lead to a misallocation of resources between economies.

Freedom to achieve domestic policy objectives

It is sometimes argued that when the exchange rate is freely floating, balance of payments surpluses and deficits cease to be a policy problem for the government, as it is then free to pursue the domestic economic objectives of full employment and growth. Market forces 'look after' the current account of the balance of payments, leaving governments free to concentrate on domestic economic policy. If, in the pursuit of domestic objectives, the inflation rate is higher than in other countries, in a freely floating world the exchange rate simply falls to restore competitiveness.

SYNOPTIC LINK

The external value of the currency is always going to affect the performance of a country's economy. Go back to Chapter 9 and refresh your understanding of the main objectives of government policy and economic indicators. Changes in the exchange rate may affect the main economic indicators and the ability to achieve the main macroeconomic objectives.

Making it easier to control inflation

A government should benefit from a floating exchange rate because the exchange rate insulates the country against 'importing inflation' from the rest of the world. If inflation rates are higher in the rest of the world, a fixed exchange rate causes a country to import inflation through the rising prices of goods imported from high-inflation countries. By contrast, a floating exchange rate appreciates, which lowers the prices of imports, insulating the economy against importing inflation.

Ability to pursue an independent monetary policy

With a floating exchange rate, monetary policy can be used solely to achieve domestic policy objectives, such as the control of inflation. This is called an *independent* monetary policy, or 'monetary sovereignty'. With a fixed exchange rate, as we explain later, interest rates may be determined by events in the outside world (and, in particular, by capital flows out of and into currencies), rather than by the needs of the domestic economy. To maintain a fixed exchange rate, interest rates may have to be raised to prevent the exchange rate from falling. In this situation, monetary policy is no longer independent, in the sense that it can no longer be assigned to pursuing purely domestic policy objectives. In a fixed exchange rate system, domestic policy objectives are often sacrificed to the pursuit of the external policy objective of maintaining an exchange rate target.

TEST YOURSELF 15.4

Explain how a fall in the value of a country's currency can help the country to correct a trade deficit.

SYNOPTIC LINK

Go back to Chapter 11, section 11.3, and read over how inflation is measured in price indices. When the value of sterling appreciates or depreciates, this impacts on the rate of inflation because the price of imports changes in the inflation price indices.

Disadvantages of floating exchange rates

Freely floating exchange rates nevertheless have some disadvantages, particularly relating to the fact that, in the modern globalised world in which financial capital is internationally mobile, capital flows rather than exports and imports can be the main determinants of exchange rates. The main disadvantages of floating exchange rates are described below.

The adverse effects of speculation and capital flows

The argument that a freely floating exchange rate is never over-valued or under-valued for very long depends crucially upon the main assumption of the traditional theory of exchange rates, that currencies are bought and sold on foreign exchange markets only to finance trade. This assumption means that speculation and capital flows have no influence on exchange rates. But this is at odds with how the modern globalised economy works. These days, well over 90% of currency transactions taking place on foreign exchange markets stem from capital flows and from the decisions of individuals, business corporations, financial institutions and even governments to switch wealth portfolios between different currencies.

In the short run, exchange rates are extremely vulnerable to speculative capital or 'hot money' movements into or out of currencies. Just like a fixed exchange rate, a floating exchange rate can be over-valued or under-valued, which means it does not reflect the trading competitiveness of the country's goods and services. In this situation, speculators will buy currencies they perceive to be under-valued in order to make capital gains when selling the currencies at higher exchange rates in the future. Likewise, they will sell currencies they perceive to be over-valued in order to avoid capital losses when selling the currencies at lower exchange rates in the future.

(Arguably, however, fixed exchange rates are even more vulnerable than floating exchange rates to speculative capital flows. This is because fixed exchange rates often become over-valued or under-valued. In the case of an over-valued currency, speculators may sell the currency in order to make a profit by buying it back at a lower exchange rate after the currency has been devalued. However, if devaluation does not take place, the speculators can instead enjoy the 'one-way option' of buying back the currency at the exchange rate at which it was sold.) This one-way option was instrumental in the pound being forced out of the European Exchange Rate Mechanism (ERM) — a precursor to the single currency — in 1992.

International trading uncertainty

It is sometimes argued that, whereas fixed exchange rates create conditions of certainty and stability in which international trade can prosper and grow, the volatility and instability caused by floating exchange rates slows the growth of, and even destroys, international trade. The prospect of changes in the exchange rate means that companies cannot be sure how much domestic currency they will have to pay for imports or receive from exports, and this uncertainty can deter them from engaging in international trade. By inhibiting trade, resources may be misallocated and world living standards reduced.

(In reality, however, *hedging*, which involves the purchase or sale of a currency in the 'forward' market, say three months in advance of the actual delivery of the currency and payment for trade, reduces the trading uncertainties associated with floating exchange rates. Indeed, fixed and managed exchange rates may also cause uncertainty, especially when a currency is over- or under-valued and a devaluation or revaluation is expected.)

SYNOPTIC LINK

Speculation and capital flows are recorded in the financial account of the balance of payments, which we explained in Chapter 14, section 14.3 — but here we focus on how speculative capital flows between currencies affect the demand for, and supply of, currencies.

TEST YOURSELF 15.5

What is meant by 'hot money'?

Floating exchange rates may cause cost-push inflation

Floating exchange rates sometimes contribute to cost-push inflation. Suppose a country has a higher rate of inflation than its trading partners:

- Trading competitiveness and the current account of the balance of payments both worsen, causing the exchange rate to fall in order to restore competitiveness.
- This may trigger a vicious cumulative downward spiral of faster inflation and exchange rate depreciation.
- The falling exchange rate increases import prices, which raise the rate of domestic cost-push inflation.
- Workers react by demanding pay rises to restore the real value of the eroded real wage.
- Increased inflation itself erodes the export competitiveness initially won by the fall of the exchange rate.
- This triggers a further fall in the exchange rate to recover the lost advantage, and so the process continues.
- The resulting downward spiral can eventually destabilise large parts of the domestic economy, causing unemployment and reducing economic growth.

Floating exchange rates may also cause demand-pull inflation

Floating exchange rates can trigger demand-pull inflation as well as cost-push inflation. With a floating exchange rate, there is no need to deflate the domestic economy to deal with a balance of payments deficit on the current account.

Suppose a large number of countries with floating exchange rates simultaneously increase aggregate demand. This can lead to excess demand on a worldwide scale, which fuels global inflation. This happened in the 1970s, when a worldwide expansion of demand created conditions in which oil and primary goods producers could raise prices and still sell in world markets. In countries such as the UK, the resulting inflation appeared to be import cost-push inflation, caused by the rising cost of imported energy and raw materials.

However, the true cause lay in excess demand created by the simultaneous effects of demand expansion and floating exchange rates, when world supply could not increase, at least in the short run, to meet the surge in global demand. Similarly, floating exchange rates allow an individual government to over-inflate its economy, provided it is willing to allow the exchange rate to fall.

In contrast, the commitment to a fixed exchange rate prevents a government unilaterally adopting an excessively expansionary monetary policy, since the resulting inflation will cause the balance of payments to deteriorate and put downward pressure on the exchange rate. To prevent this, interest rates will have to be increased, damping down the domestic inflationary pressures.

TEST YOURSELF 15.6
Use *AD/AS* diagrams to distinguish between cost-push and demand-pull inflation.

SYNOPTIC LINK
Refer back to the coverage of speculative 'hot money' capital flows in Chapter 14, section 14.3, as these are very important in determining the value of a currency in the foreign exchange market.

STUDY TIP
Look up the relationship between changes in the value of sterling and the rate of UK inflation. Sterling has experienced at least two significant depreciations in recent years: 2008 and 2016. In the period after each depreciation, the rate of inflation increased. Learn the value of sterling before and after each depreciation.

TEST YOURSELF 15.7
All other things being equal, which one of the following is the most likely effect on the UK economy of a change in the exchange rate from $1 = £1.50 to $1 = £2?

A A fall in aggregate demand and a rise in inflationary pressure

B A rise in aggregate demand and a rise in inflationary pressure

C A rise in aggregate demand and a fall in inflationary pressure

D A fall in aggregate demand and a fall in inflationary pressure

Fixed exchange rates

KEY TERM
fixed exchange rate an exchange rate fixed at a certain value and maintained at this value by the central bank's intervention in the foreign exchange market.

With a freely floating exchange rate system, a currency's external value rises or falls to eliminate a balance of payments surplus or deficit. By contrast, with rigidly **fixed exchange rates**, a currency's external value remains unchanged, while the *internal* price level, or more usually the level of domestic economic activity and output, adjusts to eliminate a balance of payments disequilibrium on current account.

As the extension material on pages 583–84 explains, there are few examples of fixed exchange rates today. Note also that so-called fixed exchange rates have rarely, if ever, been rigidly fixed. Modern fixed exchange rates are better thought of as adjustable peg exchange rates. We will explain this term shortly.

STUDY TIP
A rigidly fixed exchange rate is an extreme form of a managed exchange rate.

Key features of a fixed exchange rate

Figure 15.4 illustrates the key features of a fixed exchange rate. In the first place, the country's government or its central bank announces that the exchange rate is being fixed at a particular rate and the central bank is then given responsibility for maintaining that rate. In the context of Figure 15.4, the Bank of England would announce that the pound's exchange rate is being

fixed at US$1.30. $1.30 would be called the exchange rate's 'central peg', also known as the 'central rate', 'parity' or 'par value'.

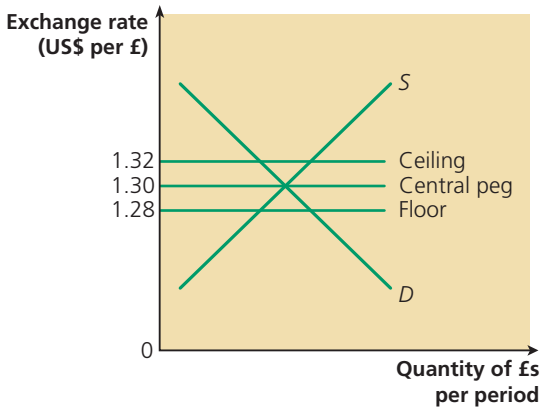


Figure 15.4 The ceiling, central peg and floor of a fixed exchange rate

Simultaneously, the Bank of England would announce 'ceiling' and 'floor' limits, set respectively just above and below the 'central peg'. Market forces (supply and demand on foreign exchange markets) then determine the currency's 'day-to-day' exchange rate.

Provided that the exchange rate remains inside the 'zone' or 'band of flexibility' between the ceiling and the floor, central bank intervention is not needed. However, if either the demand curve for, or the supply curve of, the currency shifts to a new position, in the absence of Bank of England intervention, market forces might take the exchange rate above $1.32 or below $1.28. If this happens, the fixed exchange rate has effectively broken down. To prevent this happening, whenever the ceiling or floor looks as if it is going to be breached, the central bank intervenes.

Advantages of fixed exchange rates

Because the advantages and disadvantages of fixed exchange rates are closely related to but mirror those of floating rates, which we have already explained in some depth, we will provide only a brief summary here.

The principal advantages of fixed exchange rates are:

- Fixed exchange rates attempt, not always successfully, to achieve certainty and stability in foreign exchange markets.
- Fixed exchange rates impose an anti-inflationary discipline on a country's domestic economic management and on the behaviour of its workers and firms.

Disadvantages of fixed exchange rates

By contrast, the principal disadvantages of fixed exchange rates are:

- Fixed exchange rates may *increase uncertainty* rather than *create certainty*. This is likely if a devaluation or revaluation is expected to occur in the near future.
- If the fixed exchange rate is significantly *over-valued*, it may impose the severe deflationary costs of lost output and unemployment, unless corrective action is taken to *devalue*.
- If the fixed exchange rate is significantly *under-valued*, inflation may be imported from the rest of the world, unless corrective action is taken to *revalue*.
- An independent monetary policy cannot be implemented.
- There may be recurrent balance of payments or currency crises in a country whose currency is over-valued.
- Resources may be tied up in official reserves (needed to support the fixed exchange rate), which could be used more productively elsewhere.
- Both over-valuation and under-valuation of the currency can lead to a misallocation of resources.

How governments can intervene to influence the exchange rate

These days, it is the country's central bank rather than its government that influences the exchange rate, though in a country such as the UK, the central

bank (the Bank of England) takes strong notice of the wishes of the chancellor of the exchequer. We will explain first how the central bank can try to maintain a fixed exchange rate, before extending the explanation to adjustable peg exchange rates in the extension material.

There are two main ways in which a central bank can intervene in markets to maintain a fixed exchange rate. The first is by buying or selling its own currency on the foreign exchange market. In Figure 15.4, if the exchange rate threatens to rise above the $1.32 ceiling, the Bank of England will sell its own currency on the foreign exchange market to offset excess demand for pounds that would otherwise pull up the exchange rate. Likewise, if the exchange rate is deemed likely to fall below its $1.28 floor, the Bank of England will step into the foreign exchange market and buy its own currency. In this case, by creating an 'artificial' demand for the pound, the Bank attempts to prevent an excess supply of pounds from pushing the exchange rate below the floor. The policy of buying and selling currencies to support an exchange rate is known as *exchange equalisation*.

The second method of supporting a fixed exchange rate is, in the UK's case, by the Bank of England raising or lowering Bank Rate to keep the pound's exchange rate between a ceiling of $1.32 and a floor of $1.28. Lower domestic interest rates might cause international holders of the pound to sell pounds and buy other currencies. This would cause the pound's exchange rate to fall, hopefully to a rate below the ceiling of $1.32. Conversely, an increase in UK interest rates attracts 'hot money' into the pound. Through the selling of other currencies and the buying of the pound, short-run capital inflows raise the demand for pounds, hopefully taking the pound's exchange rate above its floor rate of $1.28.

In real life, both methods of intervention are likely to be used together to support each other, particularly if an exchange rate crisis has led to a flow of money out of the economy which threatens to destroy the fixed exchange rate. Perhaps the most important point to note is that, when supporting a fixed exchange rate, a country's central bank may not be able to change its key interest rate (Bank Rate in the UK) to achieve purely domestic macroeconomic policy objectives such as economic growth and low unemployment.

Instead, the freedom to use monetary policy to achieve domestic objectives is sacrificed on the altar of supporting a fixed exchange rate. This fact of life probably explains why fixed exchange rates are rare today and why floating exchange rates are the norm. In 2023, however, there were still around 30 countries with formally or informally fixed exchange rates — most were fixed against either the euro (e.g. Denmark) or the US dollar (in Latin America).

SYNOPTIC LINK

If a government pursues a high exchange rate policy to combat inflation, this will have significant implications for the setting of monetary policy and the work of the central bank. Go back to Chapter 12 and read over section 12.3.

EXTENSION MATERIAL

Managed exchange rates

A rigidly fixed exchange rate is, of course, an example (the extreme example) of a managed exchange rate. However, there are two other examples of managed exchange rates: *adjustable peg* and *managed floating* or *'dirty' floating*. In both cases, a country's monetary authorities hope to achieve the stability and certainty associated with fixed exchange rates combined with a floating exchange rate's ability to avoid over-valuation and under-valuation by responding to market forces.

Adjustable peg exchange rates

An adjustable peg exchange rate is similar to a rigidly fixed exchange rate, except that the central bank may alter the exchange rate's central peg by devaluing or revaluing. If you look back at Figure 15.1, you will see that on the spectrum of exchange rates, an adjustable peg exchange rate is much closer to a rigidly fixed exchange rate than to a freely floating exchange rate. By contrast, a 'dirty-floating' (or managed-floating) exchange rate is much closer to a freely floating exchange rate.

Nevertheless, adjustable pegs are more flexible than the central pegs of rigidly fixed exchange rates, such as the old nineteenth-century gold standard. This is because the exchange rate can be adjusted upwards or downwards from time to time by the country's central bank. Adjustments take place because, over a period of time, exchange rates are likely to become over- or under-valued as, for example, countries experience different rates of inflation. An upward revaluation corrects an under-valued exchange rate, whereas a downward adjustment or devaluation is used to correct over-valuation.

Figure 15.5, which is a development of Figure 15.4, illustrates devaluation of the pound in an adjustable peg exchange rate system. The exchange rate is initially fixed at a central peg of $1.30. As we explained earlier, the interaction of supply and demand for the currency on the foreign exchange market then determines the day-to-day exchange rate. Provided the exchange rate stays between a ceiling and a floor set at the time that the central peg was fixed, there is no need for central bank intervention. However, the graph shows the sterling exchange rate falling to the floor of $1.28, possibly because of a speculative capital flow against the currency. At this point the Bank of England intervenes, raising domestic interest rates to attract capital flows into the currency and purchasing the pound on the foreign exchange market using its foreign exchange reserves. By selling reserves and buying its own currency, the central bank creates an artificial demand for the pound, bidding up the exchange rate.

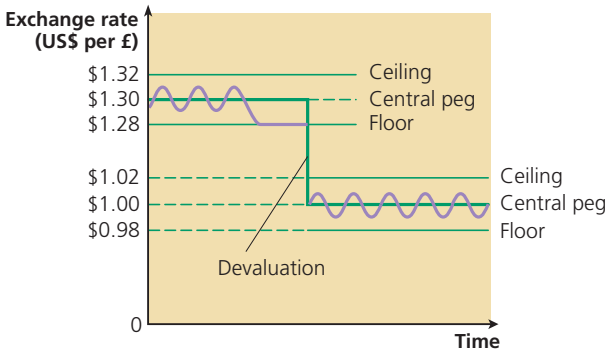


Figure 15.5 A devaluation of an adjustable peg exchange rate

Persistent support for a currency almost certainly means that its exchange rate is over-valued, condemning the country to over-priced exports, under-priced imports and a persistent current account deficit. With the rigidly fixed exchange rate shown in Figure 15.4 this is the end of the story. The country's government has to deflate the domestic economy and/or impose import controls, since devaluation and revaluation are not permitted in a rigidly fixed system.

However, Figure 15.5 shows the Bank of England devaluing the exchange rate to a new central peg of $1.00 in an attempt to correct the fundamental imbalance. This illustrates the difference between adjustable peg and rigidly fixed exchange rate systems. But if a country makes frequent changes to the central peg, the benefits of having a fixed exchange rate are lost.

TEST YOURSELF 15.8
Suggest three reasons why a government might seek to manage its exchange rate.

Managed-floating or dirty-floating exchange rates

Fixed and adjustable peg exchange rate systems have now been abandoned throughout much of the world. Most exchange rates now float, though there is a difference between 'clean floating' and 'dirty floating'. Clean floating is the same as freely floating and pure floating — that is, there is no central bank

intervention to prop up the exchange rate or to manipulate its value. In contrast to clean floating, 'managed floating' or 'dirty floating' occurs when the exchange rate is officially floating, in the sense that a country's monetary authorities announce that market forces are determining the exchange rate, but in fact they intervene unofficially behind the scenes, buying or selling their own currency in order to influence the exchange rate. (Countries around the world admit that they 'manage' their floating exchange rates. Arguably there are currently no countries that have a truly free-floating exchange rate system, with no central bank intervention at all.)

At the time of writing (in July 2022), it is probably true to say that for most of the period since September 1992, when the pound left an adjustable peg exchange rate system within the European Union, the pound's exchange rate has more or less freely floated. Nevertheless, the Bank of England does intervene on a daily basis, as part of a 'smoothing operation' through which it tries to iron out temporary fluctuations in the pound's exchange rate. In recent years, the UK has experienced two major currency crises. In the months after the 2008 financial crisis, sterling lost approximately 25% of its value against the US dollar; and after the Brexit referendum on 23 June 2016, the pound lost approximately 15% of its value against the US dollar.



The Bank of England intervenes frequently to manage the sterling exchange rate

STUDY TIP

Research the performance of the pound sterling since 2000 against the US dollar and the euro. Look carefully at how the value of the pound fluctuated throughout the period and examine the value of the currency in the months following the depreciations in 2008 and 2016.

APPLICATION OF ECONOMICS IN THE REAL WORLD 15.1

Devaluation of the pound and UK economic performance

The devaluation of a nation's currency can be used by a government to reduce a current account deficit on the balance of payments and stimulate economic growth. In recent decades, three significant devaluations or depreciations of the pound have taken place. On each occasion, sterling lost almost a quarter of its value against the US dollar over a six-month period. Each devaluation or depreciation took place for very different reasons and the impact on UK economic performance has varied.

The pound's ejection from the Exchange Rate Mechanism in 1992

The Exchange Rate Mechanism (ERM) was a system of adjustable peg exchange rates set up by the European Union in 1979, within a newly created European Monetary System (EMS), in preparation for economic union within the EU. Having initially opposed membership of the ERM, the UK joined the system in 1990, but at a substantially over-valued exchange rate. On 16 September 1992 the UK government was forced to abandon its economic policy and withdraw the pound from the ERM, a currency system which was meant to stabilise the exchange rates of member countries.

When the pound joined the ERM in 1990, the 'high-pound' objective was meant to bleed inflation out of the economy by forcing UK firms to make efficiency savings to stay price competitive in world markets. However, this harmed the UK economy in two painful ways. First, the strong currency made imports cheap and exports expensive. Although this helped bring down the UK's rate of inflation, UK firms struggled to survive in world markets.

Second, the Bank of England raised interest rates to attract 'hot money' flows into the currency and

maintain the high exchange rate. However, this put UK households with high levels of mortgage debt under tremendous pressure and restricted the level of consumer spending.

When, in 1992, the markets forced the UK government to devalue the pound, the UK economy was in a deep recession, which itself had been caused by the pound's over-valuation. At the time, the forced devaluation of the pound was seen as a national humiliation, with the result that the day on which it took place was nicknamed 'Black Wednesday'. Many link the rising influence of 'Euro-scepticism' to this era.

However, immediately after the devaluation, the economy started to grow. The low exchange rate made UK exports price competitive and in the following years the UK's unemployment rate fell and the current account deficit on the balance of payments was eliminated. Between 1992 and 1997, ONS statistics show that exports increased by 53% and the economy grew by 15.2%.

The departure from the ERM also had significant implications for UK monetary policy. The Bank of England no longer needed to maintain high interest rates to support a fixed exchange rate, so falling interest rates significantly reduced the cost of borrowing for both firms and households. Looser monetary policy created the conditions for increased consumer spending and business investment.

The Great Recession, 2008 and 2009

In the summer of 2008, the dollar exchange rate was £1 = $2.00. Then the banking crisis shook the world. The UK's banking system was at the centre of the worst economic crisis since the 1930s and the value of the pound crashed as 'hot money' flowed out of the pound and searched for safety in other currencies. By January 2009, the exchange rate had fallen to £1 = $1.30.

To stimulate the domestic economy, the Bank of England slashed Bank Rate to 0.5%, and in 2009 it started the unorthodox monetary policy of quantitative easing. The fall in the exchange rate, record low interest rates and the injection of billions of pounds of electronic money into the economy should, according to conventional economic theory, have stimulated the economy.

However, it took seven years for the UK economy to recover from the recession. After 2009, economic growth was weak and UK firms failed to experience an increase in demand in their export markets. The size of the deficit on the current account remained high, only reducing to 2% of GDP in 2011, two years after the devaluation. Fiscal austerity at home and the eurozone crisis abroad were major reasons why the UK recovery was so weak. However, the devaluation of sterling did eventually stimulate the economy in the same way that it had after earlier devaluations.

After the Brexit referendum in 2016

Following the publication of the 2016 Brexit referendum result, the pound's exchange rate fell against both the dollar and the euro. In January 2016, the exchange rate was £1 = $1.45. A year later, in January 2017, it had fallen to £1 = $1.20. The devaluation had two predictable outcomes: cost-push inflation squeezed living standards. The CPI rate of inflation in the two-year period after the devaluation increased to over 3% a year, but the current account deficit was little changed.

External shocks and the fall in the value of the pound

Many financial investors view the US dollar as a 'safe' investment — it is still seen as the global 'reserve currency' that can be held as a safe place to store financial capital (i.e. the dollar generally holds its value and is highly unlikely to lose value over a short space of time).

As a result of this, when there is a global external shock, the US dollar generally rises as investors look for a 'safe' currency for any short-term capital holdings. This increase in the value of the US dollar automatically means other currencies will lose value against the dollar. In early 2020 as news of Covid-19 first appeared and in 2022 due to the crisis in Ukraine, the pound lost about 10% of its value against the dollar (it did recover the lost ground on the first occasion, but at the time of writing it has not yet recovered the loss in value after the second shock).

Some conclusions

The devaluations of 2008 and 2016 affected UK economic performance in a different way from what had happened after the earlier 1992 devaluation. In 1992, export competitiveness rapidly improved, export-led growth occurred, and the economy recovered quickly. By contrast, the falls in the value of the pound in the later cases did not lead to strong export-led growth or to the elimination of the current account deficit.

This difference was partly because the nature of global capitalism in the twenty-first century has

changed. Modern global supply chains now require that the goods and services which UK firms export are priced in US dollars and euros, and not in UK pounds.

As a result, a fall in the value of the pound has a limited effect on the price competitiveness of UK exports. The 2008 and 2016 devaluations or depreciations of the pound did not significantly boost exports or decrease imports. Hence, the evidence appears to show that these devaluations brought limited gains for UK exporters, while inflicting painful imported inflation on UK households.

Follow-up questions

1 Explain how a central bank will implement monetary policy to try to maintain a high exchange rate.
2 Explain how a depreciation in the value of a currency affects the rate of domestic inflation.
3 Suggest why a fall in the pound's exchange rate has not in recent years led to a significant increase in demand for UK exports.
4 Describe what happened to the pound's exchange rate during the year before you are reading this case study.

TEST YOURSELF 15.9
How does exchange rate stability affect businesses planning long-term investment projects?

STUDY TIP
Research the relationship between the Danish krone and the euro. Look carefully at the exchange rate of the currencies since 1999. The Danish krone has fluctuated throughout the period, but its movements have been very minor compared to that of, for example, the UK pound against the euro.

Currency unions, the euro and the eurozone

Currency unions

KEY TERM
currency union an agreement between a group of countries to share a common currency, and usually to have a single monetary and foreign exchange rate policy.

The United Kingdom comprises a number of countries, but all of them use the pound as a common currency. Likewise, the 50 states of the United States of America, plus the federal district of Washington DC, all use the US dollar. Both the UK and the USA are examples of a **currency union**, which is a group of countries using or sharing a single currency. A currency union is often known as a monetary union.

In the 1990s, the UK, then being a member state of the European Union (EU), was given the choice of keeping its own currency (still within the EU), or of adopting the common European currency that at the time was being set up. If the UK had chosen the latter, it would have replaced the pound with the new common currency, the *euro*, and joined what became known as the *eurozone*.

TEST YOURSELF 15.10
Explain how a country that joins a monetary union loses control of its monetary policy.

The euro and the eurozone

The euro came into existence on 1 January 2002 when euro notes and coins entered circulation. However, economists usually date the euro's introduction as 1 January 1999, as this was the date on which the exchange rates of the 12 countries that became the first members of the eurozone (or euro area) were irrevocably fixed against each other.

In the early 1990s, the Treaty of the European Union (or Maastricht Treaty) had begun the process which eventually created monetary union within the eurozone. At the time, monetary union was considered necessary for two reasons. In the first place, although the Single European Act had eventually created almost completely free internal trade in goods and services, the need to exchange national currencies in this trade increased transaction costs. A single currency was thus deemed necessary to complete the single market.

Secondly, upward and downward movements of EU national currencies led to under-valued and over-valued exchange rates within the trading bloc. Exchange rate fluctuations artificially increased the competitiveness of those countries benefiting from a falling exchange rate, while reducing the competitiveness of countries whose exchange rates were rising. Monetary union would create a level playing field for all member countries.

The UK government decided to keep the pound and not to replace it with the euro, which means that along with nine other EU member currencies, the pound was not in the eurozone. At the time of writing in February 2023, 20 EU member states were in the eurozone, together with non-member states such as Kosovo and Montenegro. Additionally, various other countries pegged their currencies against the euro.

Currency unions and economic integration

The euro was created to facilitate greater economic integration among EU member states. Indeed, the euro is regarded by many as a stepping stone to full monetary union between EU states and, possibly in the future, to a much fuller economic and monetary union (EMU). At this point it is worth noting that EMU can mean two different things. The official EU meaning is *economic and monetary union*. Defined in this way, EMU suggests that common monetary arrangements adopted by EU member countries are part of a grander scheme to integrate the national economies of member states. More narrowly defined, the acronym means *European monetary union*, which involves a common monetary policy applied to all EU member states adopting the euro. In the latter meaning, EMU can be interpreted simply as a step towards making the EU's single market work better and more efficiently.

The impact of the euro on eurozone economies

When the eurozone first came into existence, it was claimed that member countries would benefit in the following ways: reduced transaction costs, elimination of currency risk, greater transparency and possibly greater competition because prices are easier to compare. According to Paul Krugman (see also Application of Economics in the Real World 14.3):

> The creation of the euro was supposed to be another triumphant step in the European project, in which economic integration has been used to foster political integration and peace; a common currency, so the thinking went, would bind the continent even more closely together.

But Krugman then stated:

> What has happened instead is a nightmare: the euro has become an economic trap, and Europe a nest of squabbling nations. Even the continent's democratic achievements seem under threat, as dire economic conditions create a favourable environment for political extremism. Who could have seen such a thing coming?

Mobility of labour

Krugman and many other economists did foresee chaos in some of the countries which use the euro because they realised that the eurozone is *not* an *optimal currency area* (see the extension material on page 589). The USA and the UK, which we mentioned earlier as common currency areas, are much closer to being optimal currency areas. Each uses a single currency, enjoys

complete internal free trade, and has a fiscal policy as well as a monetary policy covering the whole of the union. The USA, in particular, also enjoys high mobility of labour which enables workers who lose their jobs in the poorer parts of the country to find jobs in richer states such as California.

Mobility of labour and a common fiscal policy are essential if a common currency area is to become an optimal currency area. Without outward labour mobility from poorer parts of the union to the richer parts or perhaps to other parts of the world, the main way a country can regain jobs that have been lost by lack of competitiveness is through a large fall in real wages to make the region more competitive. Without a single currency, a country could restore its competitiveness by devaluing its currency. But in a currency union this is impossible. A high degree of labour mobility is therefore needed to deal with the unemployment problem. Emigration can shrink the size of the labour force to match the number of jobs available. However, it is often the young and more dynamic workers who move and this can be counterproductive.

SYNOPTIC LINK

Go back to Chapter 11, section 11.2, and refresh your understanding of unemployment and the problem of labour immobility. The potential labour immobility between eurozone member countries was seen as a weakness in the likely viability of the currency union. It may seem obvious why labour is less mobile in the eurozone compared with the 'currency union' of the USA; different languages, different laws, and different cultures and customs all make it less easy for labour to move freely around the eurozone, despite there being freedom of movement as part of being in the EU.

Importance of a common fiscal policy

A common fiscal policy is also particularly important, since it allows wealth to be transferred by a centralised fiscal authority from the richer to the poorer parts of the union. Transfers of wealth, via taxation and public spending, are essential to counter the fact that poorer parts of the union cannot achieve competitive advantage by devaluing their national currencies. Fiscal transfers can help to reduce inequality but can also be used to try to improve the competitiveness of the poorer regions: for example, by spending on improving infrastructure.

An optimal currency area thus requires coordination in fiscal policy as well as in monetary policy. The European Central Bank (ECB) achieves monetary policy coordination, but as Greece found to its expense, there is no supranational authority in the eurozone similar to the ECB to coordinate fiscal policies and to facilitate significant fiscal transfers.

The lack of fiscal policy coordination has led to some eurozone countries having high levels of government debt. This has built up especially in countries such as Greece and Portugal, in the southern eurozone. Governments in this part of the eurozone believed they would always be helped by the richer governments of countries such as Germany and the Netherlands. Also, prior to the 2008–09 recession, these southern eurozone countries could borrow cheaply on international capital markets. During that time, any euro-denominated debt was pretty much acceptable to investors, no matter which eurozone country issued the debt.

SYNOPTIC LINK

Fiscal transfers occur when a government uses the taxation and benefits system to transfer income and wealth from one section of society to another. This was covered in Chapter 7. In this chapter we are considering potential transfers on a eurozone-wide basis.

Conclusion: an incomplete single market

The single market means that in theory there is complete mobility of both labour and capital within the EU. In practice, however, labour mobility is limited, particularly in comparison with the USA where workers move in large numbers from state to state. With regard to capital, before the financial crisis in 2007 there was massive capital movement from Europe's core — mainly Germany — to its periphery, leading to an economic boom in the periphery and significantly higher inflation rates in the southern eurozone countries than in Germany. But when private capital flows from the core to the periphery came to a sudden stop in 2008, the southern countries were uncompetitive and left with prices and unit labour costs well out of line with those in the core. Suddenly the eurozone faced a major adjustment problem that it has yet to deal with fully. (So-called austerity measures appear to have had some success in countries such as Ireland, but less so in Greece.)

SYNOPTIC LINK

In Chapter 14, section 14.2, we looked at trade blocs. Read over this section and ensure you can explain the difference between a customs union and a single market, and why the EU has an incomplete single market. The closer a trading block is to a single market, the more it will behave like a competitive market.

EXTENSION MATERIAL

Optimal currency areas

If several countries use the same currency, their inhabitants will benefit from efficiency gains. They no longer have to worry about possible future changes in exchange rates, and the costs involved in currency conversion. These gains will be greater the more economically interconnected the countries are, in terms of their trading relationship, the freedom of labour and capital to move between countries, and a unified fiscal policy covering all member countries. In these circumstances, the advantages of adopting a common currency can lead to benefits for all the member states within the currency union.

The 50 states of the USA satisfy most or all the requirements for benefiting from a common currency, namely a large amount of trade in goods and services, free movement of factors of production across the union, and a unified government budget with fiscal transfers from the richer to the poorer states.

The European countries that have adopted the euro as their common currency do not satisfy those requirements. Many economists believe that the eurozone countries adopted the single currency for political rather than economic reasons. Although the eurozone countries trade a lot with each other, capital and especially labour are not sufficiently mobile between countries, and the countries lack a unified 'government' budget. Even with freedom of movement allowed under EU law, workers are less mobile (i.e. less willing to physically move for jobs) across borders in the eurozone.

The point about the lack of a unified 'government' budget is really significant. A common fiscal policy provides a mechanism for transferring resources — for example, in the form of subsidies or lower tax rates — from the more prosperous parts of the currency union to the less prosperous countries. This is what happens in the UK, for example, when taxes paid by Londoners are directed into public spending projects in Scotland, Wales and Northern Ireland. The UK and the USA are successful common currency areas, but the eurozone is much less successful, since it lacks a common fiscal policy.

SECTION 15.1 SUMMARY

- An exchange rate is the external price of a currency in terms of another currency.
- There are three main types of exchange rate: freely floating, fixed and managed floating.
- A freely floating (cleanly floating) exchange rate is determined by supply and demand.
- Provided there are no capital flows, a freely floating exchange rate should automatically eliminate a trade deficit or surplus.
- Free floating enables a country to pursue an independent monetary policy.
- There are, however, disadvantages with a freely floating exchange rate system, such as exchange rate instability or volatility.
- Speculative or 'hot money' capital flows also destabilise floating exchange rates, sometimes leading to severe over-valuation or under-valuation.
- In a rigidly fixed exchange rate system, the exchange rate cannot rise or fall to eliminate or reduce a payments surplus or deficit.
- The advantages of a fixed exchange rate relate closely to the disadvantages of a floating rate, while the disadvantages similarly relate to a floating system's advantages.
- Adjustable peg and 'dirty floating' are types of managed exchange rates.
- In an adjustable peg system, the exchange rate can be revalued or devalued to try to correct a fundamental payments imbalance.
- The euro has replaced national currencies for the majority of EU member states that are in the eurozone.
- In recent years there has been a eurozone crisis, partly caused by the fact that the eurozone is not an optimal currency area.

15.2 Economic growth and development

The difference between growth and development

SYNOPTIC LINK

Refer back to pages 366–74 in Chapter 11 for our first mention of economic growth and development. In that chapter, we were looking mainly at growth in a developed economy, such as the UK. Here, we are focusing more on growth in less developed economies.

In Chapter 11, we explained the concept of economic growth and mentioned the difference between economic growth and economic development. There we said that *economic growth* measures changes in the physical quantity of goods and services that an economy actually produces, or has the potential to produce, but that *economic development* goes further, encompassing not just the increase in *quantity* of output, but also its *quality* and contribution to human happiness.

Economic growth does not necessarily improve the economic welfare of all or most of the people living in a country. Consider, for example, the situation when the ruling elite in a country use the fruits of growth to buy military and police equipment which they then use to suppress and possibly kill those in the population who wish to change how the economy operates. Although economic growth has occurred, economic development has not.

As stated in Chapter 11, economic development can be measured by:

- a general improvement in living standards which reduces poverty and human suffering

- access to resources such as food and housing that are required to satisfy basic human needs
- access to opportunities for human development (e.g. through education and training)
- sustainability and regeneration, through reducing resource depletion and degradation
- access to decent healthcare.

STUDY TIP

Don't confuse *sustainable growth* with *sustained growth*. The latter simply means achieving a particular rate of economic growth (on average) over a number of years. For example, from the end of recession in 1992 until the onset of the next recession in 2008, the UK sustained an average annual growth rate of about 2.75%. However, a growth rate which is sustained over quite a long period may nevertheless be unsustainable if it leads to environmental problems, the depletion of vital finite resources and the degradation of others.

SYNOPTIC LINK

In Chapter 14 we looked at how multinational corporations have built global production platforms and located manufacturing facilities in emerging market economies in southeast Asia. Countries such as China and India have created hundreds of millions of industrial jobs which have lifted millions of people out of poverty.

The transfer of advanced technologies and capital equipment has helped less economically developed countries industrialise, but it has also resulted in environmental destruction and high levels of pollution. Go back over Chapter 8, section 8.4, and link the concept of externalities to the topic of development.

TEST YOURSELF 15.11

Using a diagram, explain how an economy can grow in both the short run and the long run.

The main characteristics of less developed economies

The term 'less developed economies' covers a wide range of economies, from the extremely poor to what the United Nations calls higher middle-income economies. There are various ways of identifying the main characteristics of less developed economies, but we concentrate on only one, which was developed nearly 70 years ago by W.W. Rostow, an economic historian and political theorist.

- Rostow used the term 'traditional society' to describe economies that were completely lacking in economic development. He described these as usually poor societies in which very little changes from generation to generation, and in which tradition, generally accepted customs and persistent relationships between people govern economic life. Traditional societies face a low limit to their total production which, because of limited available methods of production, is largely agricultural. Today, traditional societies survive very often as tribal societies in remote areas of countries such as Brazil and India, and in parts of Africa.

- Higher up the ranking of less developed economies are what Rostow called 'traditional economies preparing for take-off'. These are economies which are becoming more productive and in which various preconditions for successful industrial growth are beginning to appear. Scope for commerce and trade appears, along with banks and other financial institutions for the mobilisation of savings into productive investment.
- At the next stage, Rostow identified economies taking off into self-sustaining growth. This involves change from a largely agricultural to an industrialised or manufacturing economy. Output increases, but only a small fraction of society benefits significantly from improved living standards. Growth takes priority over economic development.
- At the final stage, higher-income developing economies are the end result of continued self-sustained growth. At this stage of development, countries are in a position to choose to allocate the increased output that their economies are producing to social welfare and to other aspects of what we now call economic development.

Rostow's description of economic growth and development has had many critics, who generally argue that it is historically and factually wrong and is too simplistic. Nevertheless, it serves as a starting point for understanding some of the key features of less developed economies and the stages they go through to develop into prosperous societies with high levels of consumption.

TEST YOURSELF 15.12
Explain how the taxation revenues of a country which exports large volumes of oil will be affected if the price of a barrel of oil suddenly falls significantly on world markets.

SYNOPTIC LINK
Go back to Chapter 1 and read over section 1.5. Use a production possibility diagram to explain why a country in the early stages of industrialisation will focus on capital investment to achieve a faster rate of economic growth in order to achieve higher living standards in the long run.

Indicators of development

KEY TERM
indicators of development these include gross domestic product (GDP) per head, information on the distribution of income, mortality rates and health statistics.

Economists usually use gross domestic product per head or *GDP per capita* as their first **indicator of economic development**. Unfortunately, for most developing economies, GDP is usually greater than another national income indicator of development, *gross national income* or GNI. Profit outflows and interest payments out of developing economies to more developed economies, and to banks within these economies, explain why this is the case. (When national income statistics are used as indicators of development, GDP per capita or GNI per capita (where per capita means per head of population) should be used rather than 'raw' GDP or GNI statistics.)

But are measures of national income or output, whether GDP or GNI figures, the best measures of standards of living, economic welfare and economic development? To try to answer this question, it is useful first to identify three components of economic welfare:

total economic welfare = economic welfare derived from goods and services purchased in the market economy + economic welfare derived from public goods and merit goods provided collectively by the state + economic welfare derived from quality of life factors, including external benefits minus external costs and intangibles

If used carefully, national income figures can provide a reasonable estimate of economic welfare derived from the first two of these three elements, both of which relate to the direct consumption of material goods and services. However, national income fails to provide a satisfactory indication of how externalities and other quality of life factors affect economic welfare and living standards.

Intangible factors, which are the third element in people's living standards, are largely ignored. These intangible factors include the value people place on leisure time and living close to work, and the externalities such as pollution and road congestion generated from the production and consumption of national income, which affect people's welfare and quality of life.

National income also fails to reflect the impact of resource depletion and environmental degradation, resulting from producing current income, on humankind's ability to produce future income. This means that national income and GDP do not address the issue of sustainability.

In addition, while national income includes the value of the output of merit goods such as health and education, unless the data are broken down, it is not possible to determine the extent to which the nation's resources are devoted to producing these services which are vital to economic development. Indicators such as life expectancy, infant mortality rates and literacy rates can be used to supplement national income per head in order to provide a better indicator of people's quality of life.

The United Nations Human Development Index (HDI)

KEY TERM
United Nations Human Development Index an index based on life expectancy, education and per capita income indicators, which ranks the world's countries into four tiers of human development: (i) very high human development; (ii) high human development; (iii) medium human development; (iv) low human development.

The environmental pressure group Friends of the Earth has argued that measures of national income such as GDP were never intended to be indicators of a country's economic welfare or stage of development. Other indicators of development which are less dependent on 'raw' GNI or GDP are increasingly used to place a value on economic and social progress. One of the earliest of these was the *Measure of Economic Welfare* (MEW) developed by Nordhaus and Tobin in 1972. More recent attempts to adjust conventional national income figures to show developments in economic welfare include the **United Nations Human Development Index** (HDI) and the *Index of Sustainable Economic Welfare* (ISEW).

The HDI is constructed by measuring:

- life expectancy at birth
- mean years of schooling and expected years of schooling
- gross national income (GNI) per head of population, reflecting purchasing power parity (PPP) in US dollars.

The maximum value of the HDI is 1 (or unity). The closer a country's HDI is to 1, the greater is its human development, measured in terms of the three indicators specified in the index.

Table 15.2 shows the 15 highest-ranked countries in the HDI for 2022, based on 2021 data, and also the 15 lowest-ranked countries.

In 2021, Switzerland ranked first on the Human Development Index



In 191st place, South Sudan ranked last on the HDI



In 2022 Switzerland led the way in the HDI 'top 15', with the UK and the USA not making that list, in eighteenth and twenty-first positions respectively. At the other end of the spectrum, almost all the countries in the 'bottom 15' were located in sub-Saharan Africa.

The unadjusted HDI is by no means a perfect index of human development, however, since it ignores inequalities in the distribution of income. In 2010 the UN's *Human Development Report* developed the HDI by introducing a new *Inequality-adjusted Human Development Index* (IHDI). The IHDI estimates the actual level of human development (accounting for inequality) in each of the 191 countries included in the HDI. The HDI itself estimates the maximum IHDI that could be achieved if there were no inequality within countries. The greater the degree of inequality within a country, the greater the difference between the country's HDI and IHDI.

SYNOPTIC LINK

The Human Development Index was created to focus on people and their capabilities rather than solely on economic growth. Go back to Chapter 7, section 7.2, which looked at the problem of poverty, and think about the countries shown in Table 15.2. How would you apply the concepts of absolute and relative poverty to the countries in the table?

TEST YOURSELF 15.13

Explain two ways in which higher levels of education lead to higher living standards.

Table 15.2 The 'top 15' and the 'bottom 15' countries in the Human Development Index, published 2021

HDI rank	Country	Human Development Index (HDI) value 2021	Life expectancy at birth (years) 2021	Expected years of schooling (years) 2021	Mean years of schooling (years) 2021	Gross national income (GNI) per capita (2017 PPP $) 2021	GNI per capita rank minus HDI rank 2021	HDI rank 2020
Very high human development								
1	Switzerland	0.962	84.0	16.5	13.9	66,933	5	3
2	Norway	0.961	83.2	18.2	13.0	64,660	6	1
2	Iceland	0.959	82.7	19.2	13.8	55,782	11	2
4	Hong Kong, China (SAR)	0.952	85.5	17.3	12.2	62,607	6	4
5	Australia	0.951	84.5	21.1	12.7	49,238	18	5
6	Denmark	0.948	81.4	18.7	13.0	60,365	6	5
7	Sweden	0.947	83.0	19.4	12.6	54,489	9	9
8	Ireland	0.945	82.0	18.9	11.6	76,169	–3	8
9	Germany	0.942	80.6	17.0	14.1	54,534	6	7
10	Netherlands	0.941	81.7	18.7	12.6	55,979	3	10
11	Finland	0.940	82.0	19.1	12.9	49,452	11	12
12	Singapore	0.939	82.8	16.5	11.9	90,919	–10	10
13	Belgium	0.937	81.9	19.6	12.4	52,293	7	16
13	New Zealand	0.937	82.5	20.3	12.9	44,057	16	13
15	Canada	0.936	82.7	16.4	13.8 e	46,808	9	15
Low human development								
177	Guinea-Bissau	0.483	59.7	10.6	3.6	1,908	0	177
178	Liberia	0.481	60.7	10.4	5.1	1,289	7	179
179	Congo (Democratic Republic of the)	0.479	59.2	9.8	7.0	1,076	9	180
180	Afghanistan	0.478	62.0	10.3	3.0	1,824	–2	177
181	Sierra Leone	0.477	60.1	9.6	4.6	1,622	–1	181
182	Guinea	0.465	58.9	9.8	2.2	2,481	–13	182
183	Yemen	0.455	63.8	9.1	3.2	1,314	1	183
184	Burkina Faso	0.449	59.3	9.1	2.1	2,118	–8	185
185	Mozambique	0.446	59.3	10.2	3.2	1,198	2	184
186	Mali	0.428	58.9	7.4	2.3	2,133	–11	186
187	Burundi	0.426	61.7	10.7	3.1	732	4	187
188	Central African Republic	0.404	53.9	8.0	4.3	966	1	188
189	Niger	0.400	61.6	7.0	2.1	1,240	–3	189
190	Chad	0.394	52.5	8.0	2.6	1,364	–7	190
191	South Sudan	0.385	55.0	5.5	5.7	768	–1	191

APPLICATION OF ECONOMICS IN THE REAL WORLD 15.2

Why GNP figures provide an inadequate indicator of economic development

In the late 1960s, US senator Robert Kennedy described how GNP on its own cannot be a measure of human development — that is, development cannot simply be measured on the economics of consumption.

> The gross national product includes air pollution and advertising for cigarettes, and ambulances to clear our highways of carnage. It counts special locks for our doors, and jails for the people who break them. GNP includes the destruction of the redwoods and the death of Lake Superior. It grows with the production of napalm and missiles and nuclear warheads ... and if GNP includes all this, there is much that it does not comprehend. It does not allow for the health of our families, the quality of their education, or the joy of their play. It is indifferent to the decency of our factories and the safety of our streets alike. It does not include the beauty of our poetry or the strength of our marriages, or the intelligence of our public debate or the integrity of our public officials ... GNP measures neither our wit nor our courage, neither our wisdom nor our learning, neither our compassion nor our devotion to our country. It measures everything, in short, except that which makes life worthwhile; and it can tell us everything about America — except whether we are proud to be Americans.
>
> US Senator Robert Kennedy, 1967

Follow-up questions

1. Distinguish between GNP and GDP.
2. What are the three methods of measuring national income?
3. Why was the United Nations' Human Development Index (HDI) created?
4. Taking into account what you have learnt earlier in your studies about negative externalities and demerit goods, to what extent does this extract support the view that national income is of little use in measuring economic development?

Factors that affect growth and development

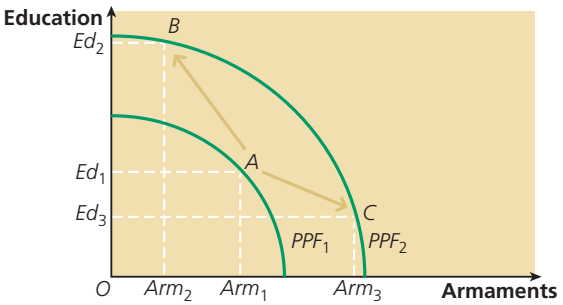


Figure 15.6 Economic growth and development in a country producing just two goods: education and arms

A large number of factors affect both growth and development. One of these that we covered in some depth in Chapter 10 is investment. Investment in new capital goods and technical progress are two of the main factors leading to long-term economic growth and they are also necessary for economic development. Consider, however, Figure 15.6.

For the simplified economy depicted in Figure 15.6, the outward movement of the economy's production possibility frontier from PPF_1 to PPF_2 illustrates economic growth. However, the movement from point A on PPF_1 to point B on PPF_2 shows that the society is now using more of its available resources to educate its children and is actually devoting fewer resources than before to arms production. The resources devoted to education increase from Ed_1 to Ed_2, whereas those devoted to armaments fall from Arm_1 to Arm_2. Increased investment in the economy has led to economic development as well as economic growth.

By contrast, a movement from point A to point C illustrates economic growth, but almost certainly not economic development, as resources have been shifted from education to armaments. Although new investment has moved the production possibility frontier from PPF_1 to PPF_2, resources have also been shifted from education into arms production. Armament production increases from Arm_1 to Arm_3, but at the expense of education, the output of which falls from Ed_1 to Ed_3. The message is that economic growth can occur without economic development, and indeed at the expense of economic development.



Education is important both for economic growth and for economic development

KEY TERM
education and training education develops individual knowledge and intellect, while training develops work skills. Both are necessary for economic growth and development.

Figure 15.6 also points to the fact that investment in **education and training** is usually an important part of both growth and economic development. Though often considered as meaning the same thing, education and training are in fact slightly different. Education furthers individual knowledge and develops a person's intellect. While a highly educated person is often more employable, however, education is not in itself about getting a job. By contrast, training is about getting a job, or about improving or diversifying skills when in a job. Training is undertaken in order to gain a specific work skill, the possession of which makes a person more employable, either in their existing line of work or through promotion or redeployment into a different type of work.

TEST YOURSELF 15.14
Explain why a country with a low savings ratio and limited access to finance is going to struggle to achieve economic growth.

Barriers to growth and development

There are, of course, many barriers which prevent or reduce economic growth, and with it the chance for further economic development. Five of these are corruption, institutional factors, infrastructure, inadequate human capital and a lack of property rights.

Corruption

KEY TERM
corruption a barrier holding back economic growth and development, especially in less developed economies.

Corruption, which can be defined as dishonest or illegal behaviour, often in the form of bribery, especially payments to powerful people such as government officials, prevents normal economic activity from taking place. Corruption and bribery divert scarce resources away from more productive uses to protect less efficient resource use. Production costs and consumer prices also rise: for example, through bribes paid by customers to corrupt officials and to employees of firms with which they wish to do business.

SYNOPTIC LINK

It is very difficult for firms to raise capital for investment if potential lenders and investors believe that they are likely to lose money because of corruption. Go back to Chapter 12 where we covered financial markets and think about how trust and transparency are important for a firm seeking to raise finance from banks or capital markets.

TEST YOURSELF 15.15

Suggest two ways in which corruption will affect market transactions and incentives in an economy.

SYNOPTIC LINK

Paying bribes to corrupt government officials adds cost to businesses. In Chapter 4, section 4.4, we looked at the firm's costs of production. Draw an average total cost curve showing the effect of corruption on a firm.

Institutional factors

KEY TERM

institutional factors examples include rules, laws, constitutions, the financial system and defined property rights.

Institutional factors, which can differ greatly between less developed and more developed economies, include countries' legal and judicial structures, public administration systems, the availability of, and access to, financial institutions such as banks and capital markets, the role, if any, of the civil service, attitudes to work, and the education and training systems.

The 'law of contract' is an important part of the legal and judicial systems of more developed economies, but contract law is much less well established in many less developed economies and indeed in eastern Europe. An efficient, policed and enforced 'law of contract' should be viewed as a necessary condition for businesses and consumers to 'do business with each other'. For example, without an enforced law of contract, firms may end up not supplying goods or services to consumers and other firms, even though the goods or services have been paid for. In this situation, business transactions may simply not take place.

SYNOPTIC LINK

The appropriate regulation of institutions and markets is necessary for a functioning economy. Go back to earlier chapters to read how UK governments have sought to regulate the UK economy: in Chapter 8, section 8.7, we covered competition policy; Chapter 8, section 8.8, looked at regulation; and finally Chapter 12, section 12.4, explained the regulation of the financial system.

Poor infrastructure

KEY TERM

infrastructure for the most part, the result of past investment in buildings, roads, bridges, power supplies, fast broadband and other fixed capital goods that are needed for the economy to operate efficiently.

Rich economies, during the course of their development over decades (and sometimes centuries), have built up an **infrastructure** of roads, railways, telephone and communication networks together with other key elements of a modern economy. Developing economies often lack such infrastructure, which is an important factor holding back their development. For example, a lack of roads raises firms' costs of production and makes it difficult or impossible for them to access raw materials and to develop their markets.

To a certain extent, as with other consumer and capital goods, developing economies can offset the disadvantage that lack of infrastructure creates, by using the fruits of technical progress to bypass old technology. A good example is provided by mobile phones. Use of these phones has become widespread in many developing economies, which has avoided the need to invest in an

infrastructure of traditional 'land-line' or terrestrial phones. An example in the future could be investment in drone technology to bypass the need to build traditional roads, as more freight deliveries could be made via drone rather than road vehicles.

SYNOPTIC LINK
Investment is the engine of long-term growth and employment. Revisit the content covered in Chapter 11 and apply the concepts to developing economies.

Investment in infrastructure is often done by public-sector investment in social capital, rather than by the private sector. One reason for this is the fact that, being very expensive to undertake, the investment may only pay for itself after a long period of years. Linked to this, much infrastructure investment creates external economies for the private sector. It may be difficult or impossible for infrastructure providers to charge prices to the private-sector firms that yield sufficient profit, and even if prices can be charged, they might deter economic development.

Lack of human capital

KEY TERM
human capital the skills, knowledge and experience possessed by the population.

Before beginning our coverage of development economics in this chapter, we had not made much reference to **human capital**. Our main reference, in Chapter 9, related the fact that investment in human capital (i.e. in people), illustrated in Figure 9.9, can be as important as investment in physical capital goods such as machinery if long-run economic growth and development are to take place. The term 'human capital' is widely used to describe people at work and their collective knowledge, skills, abilities and capacity to develop and innovate. It embodies their accumulated stock of skills and knowledge that are relevant to their employment.

Investments in both physical capital and human capital are of course *flows*, the purpose of which is to build up *stocks* of physical capital such as machines and trained labour that are needed for on-going growth and development to take place. The purpose of investment in labour training, and often investment in education in general, is to build up the stock of human capital.

TEST YOURSELF 15.16
Explain the difference between education that widens workers' skills and education that deepens workers' skills.

STUDY TIP
Make sure you understand why investment in human capital is important for *both* growth *and* development. Growth will be important for development but on its own is insufficient for development.

Lack of property rights

We have already explained the meaning and importance of *property rights* in our discussion of market failure in Chapter 8. We wrote there that property rights can be defined as the exclusive authority to determine how a resource is used. In the case of a private property right, the owner of private property such as a bar of chocolate in a sweet shop has the right in law to prevent other people from consuming the bar unless they are prepared to pay a price to the owner.

The traditional and largely undeveloped societies that we mentioned earlier in this section often lack private property rights. Most resources are communally owned. While to some this represents an idyllic paradise economy, to others it is a major cause of a lack of economic development. According to this latter view, the establishment of, and ability to trade in, private property rights is a necessary pre-condition for successful economic development. Tradable private property rights underlie the development of capitalist economies, which are arguably the most successful in achieving economic development.

In contrast, the development of the Russian economy, prior to the collapse of communism, was based to a large extent on state ownership of property rights. Although this successfully created an industrial–military complex, it also created corruption and reduced individual incentives, and it was much less successful in delivering consumer goods to the general Russian population. The absence of sufficient private property rights not only diverted resources away from true economic development, but also led to the eventual disappearance of Russian communism. Unfortunately, it also led to the transfer of previously state-owned property rights to private-sector oligarchs and other members of a new super-rich elite. Some of those who gained in this way saw their assets confiscated by various countries in 2022 as part of the package of sanctions imposed on Russia and some Russians in response to its invasion of Ukraine.

SYNOPTIC LINK

Refer back to Chapter 8 and refresh your knowledge of public goods, external costs and benefits, and road pricing.

Policies that might be adopted to promote economic growth and development

According to the OECD, sustainable development can be interpreted in economic terms as 'development that lasts' — that is, a path along which the maximisation of human well-being for today's generations does not lead to declines in future well-being.

Macro and micro policies to promote economic growth

An obvious point from which to start our analysis of appropriate policies for promoting economic growth and development is the standard macroeconomic and microeconomic government policies explained in earlier chapters. These include demand-side fiscal policy and monetary policy and supply-side policies, all of which can be focused on achieving economic growth and development and overall economic stability.

The danger is that pursuing 'growth for growth's sake' might lead to outcomes in which the costs of growth, such as environment depletion and degradation, exceed the benefits of higher living standards and economic welfare that growth is supposed to bring about. If this is the case, growth may very well be at the expense of development.

Hence the need for appropriate microeconomic policies to reduce or correct the various market failures that growth may bring. These can include the use of taxes and subsidies, respectively, to punish or deter the production of negative externalities and demerit goods, and to encourage the production of positive externalities and merit goods. Redistributive policies may also be deemed necessary to narrow the unjustifiable inequalities in the distributions of income and wealth, although, as we have seen, such policies may harm the incentives provided by markets that many think are essential for sustained growth and development. However, there is a growing consensus that a high level of inequality is a barrier to growth and many would argue that an excessive degree of inequality is incompatible with development. An equitable distribution of income is generally regarded as a characteristic of a developed economy.

Economists and politicians sometimes argue that a higher level of state intervention is required to bring about growth and development in developing economies than in more developed economies. Believers in the virtues of free markets disagree, citing the many cases of government failure that state intervention brings about. In recent decades, pro-free-market approaches have generally gained the upper hand, though not completely so. The strategic trade

theory justification for a degree of protectionism in developing economies is a case in point (see Chapter 14 for more on protectionism). In practice, there is a strong case for combining market-based instruments with some forms of government intervention.

SYNOPTIC LINK

Go back to Chapter 11 and read over the explanation of economic growth. Then read over Chapter 12, section 12.3, which covered monetary policy, and Chapter 13, which looked at fiscal policy and supply-side policy. Also, reread Application of Economics in the Real World 8.5 in Chapter 8, which looks at atmospheric pollution caused by economic growth. Think carefully about how a government develops both short-term and long-term policies to promote economic growth.

Capital flight

Some problems which affect growth and development are much more serious in developing economies than in more developed economies. *Capital flight*, which is the sudden withdrawal of money from a country, affects developing economies much more than it affects economies such as the UK. It occurs, for example, when the ruling elite within an economy move the profits they make from the country's industries into foreign bank accounts, instead of re-investing profits within the country to secure its further development. Capital flight, which is partly the result of free-market forces, can be reduced if not completely controlled by government and central bank intervention: for example, through the use of rigorously applied foreign exchange controls.

Capital flight is a major problem in sub-Saharan Africa as it impacts negatively on capital-scarce economies, such as Burundi, which generate very little saving from their own, generally very poor, populations. Trillions of dollars have flown out of several African countries over the last four decades. Capital flight has accelerated since 2000, a period that has coincided with high profits being generated from the exploitation of mineral resources in many countries in the African continent. Evidence suggests that capital flight has significantly undermined growth and development in a number of African countries, through the siphoning of potential investment funds out of the continent.

The role of aid and trade in promoting growth and development

KEY TERM

aid money, goods and services, and 'soft' loans given by the government of one country or a multilateral institution such as the World Bank to help another country. Non-government organisations (NGOs) such as Oxfam also provide aid.

Economists often debate the issue of whether **aid** is more important than trade (or vice versa) in promoting economic development. The orthodox view in countries such as the UK and the USA is that free trade and trade liberalisation are more important than aid in this respect. Free-market economists believe that international specialisation and complete free trade, undertaken in accordance with the principle of comparative advantage, benefits all the countries involved.

The strategic trade theory argument

Not all economists agree with the free-market view. The South Korean economist Ha-Joon Chang has written that 'History debunks the free trade myth' (*The Guardian*, 24 June 2002). Ha-Joon Chang's view is that governments in already

developed economies are fully in favour of free trade — but only if their countries face little or no competition from developing economies. As soon as such competition emerges, the rich countries 'pull up the drawbridge', arguing that protectionism is necessary to protect themselves from the 'unfair' competition coming from cheap-labour countries. Rich countries also argue that they need protecting from countries in the Global South which may ignore international patent laws and other aspects of intellectual copyright in order to replicate technology and products. However, transfers of technology can be an important mechanism for stimulating the development of the world's less prosperous economies.

As we explained on page 531, strategic trade theorists such as Paul Krugman have also argued that developing economies can speed up the pace of development by first selecting, and then protecting, key industries deemed vital for successful economic growth.

The counter-argument

Having said this, there is plenty of evidence that the growth of free trade and improved transport links — brought about, for example, by the containerisation of cargo — have been major factors responsible for the huge growth in international trade over recent decades. Moreover, the argument for trade rather than aid is quite broad and is not necessarily just about free trade. Global South countries can benefit from preferential trading arrangements, both with each other, and also with richer developed economies. And countering the strategic trade theory argument, trade may allow a developing economy to build up its own industries, and it encourages investment in other activities that promote development, such as good governance, property rights, infrastructure and investment in human capital.

The various forms of aid

International aid provided by richer countries and directed at less developed countries takes many forms, some of which are not really aid at all:

- **Military aid** has to be spent buying the weaponry of the donor country, although it is questionable if this can be classified as aid.
- A **'hard' loan**, usually given in a hard currency, has to be paid back at a market rate of interest, which is aid in the loosest sense.
- A **'soft' loan**, by contrast, has a below-market rate of interest attached and may even be free in the sense that, although the loan must eventually be repaid, little or no interest is charged by the donor country.
- **Disaster relief** — much of this is dispensed by non-governmental organisations (NGOs), such as the Disasters Emergency Committee (DEC), with their headquarters in developed economies. Sometimes this is matched by government-funded relief: for example, in 2021, in response to the humanitarian crisis in Afghanistan, the UK government pledged to match every pound of aid raised by the DEC, with taxpayers' money. However, disaster relief, though very necessary to limit the human cost of such events, mostly has a 'sticking plaster' effect and does little or nothing in itself to contribute to economic development. In some cases, when drought and war are ongoing for several years, as has been the case in Sudan, disaster relief may build up among sufferers a state of dependency on the generosity of those who feel honour-bound to continue to give. Several instances of corruption and sexual misdemeanours have also been reported in recent years involving NGO officials exploiting aid recipients.

- **'Tied' aid** — this can take the form of either a 'soft' loan or a gift of money that has to be spent on the exports of the country granting the aid. Arguably, this can benefit the donor country as much as, if not more than, the recipient country.
- Richer donor countries can also give aid by **'lending' their experts** to a less developed country to offer advice on matters such as improving the country's governance, its methods of production, how to use advanced technology and the maintenance of transport links and other forms of infrastructure. In recent years, China has gone one stage further by using Chinese workers and managers to build roads and other types of infrastructure in a number of African countries — though, in return, these countries have agreed to sell their exports of minerals and other raw materials to China rather than to competitor countries.

SYNOPTIC LINK
Aid is often given to a less developed country with the intention of lifting people out of poverty. Go back to Chapter 7 and consider how effective various forms of aid are at alleviating poverty.

Effective and ineffective aid

Of course, economic aid of the right type does promote economic development. For example, in a country where domestic savings are low, aid can be used to finance the domestic accumulation of capital, and it can also provide vital foreign exchange to finance the import of capital goods and resources.

However, it has sometimes been argued that the export of high-technology capital goods in foreign aid programmes, from rich to poor countries, equips countries in the Global South with the 'wrong sort of technology'. One problem facing the poor country is that when a 'high-tech' piece of equipment breaks down, the country may lack people with sufficient expertise to repair it. Additionally, spare parts produced in developed economies may have to be bought and this uses up foreign exchange, which is in short supply. Far better, it has been argued, for Global South countries to focus on intermediate, less capital-intensive technology, rather than on advanced technology. Intermediate technologies are more suited to the often cheap and plentiful labour in Global South countries, while savings needed to finance the purchase of expensive capital goods and foreign exchange are scarce.

SYNOPTIC LINK
At this point it would be useful to read again our coverage of globalisation and multinational corporations in Chapter 14. Think carefully about how the process of globalisation and the role of MNCs have affected trade and global economic activity.

The fact that both trade and aid in Global South countries are often 'skewed' to the purchase of high-technology goods often results from the desire of the countries' governments and ruling elites to invest in prestige projects such as a national airline which 'flies the flag' for the country.

Trade versus aid: a conclusion

In conclusion, free-market economists argue that trade is more important than aid in relation to economic development. Until quite recently, it was said that aid to Africa actually made the continent poorer rather than better off. This was contrasted with Asia where economic development was the result of increased trade rather than aid. However, in its 2014 report, *Economic Development in Africa*, UNCTAD, the United Nations body responsible for dealing with development issues, particularly international trade, said:

> Africa has experienced high and continuous economic growth in the past decade, prompting analysts to argue that the continent has reached a turning point in its development history and is poised to play a more significant role in the global economy in the twenty-first century. Unlike in the 1980s and 1990s, Africa's average growth rate since the turn of the millennium has also been higher than the average growth rate of the world economy.

UNCTAD believes that international aid has played a significant role in economic development across Africa, although much aid has been disaster relief rather than true economic aid. China is now providing much of the aid given to African countries, albeit with many strings attached, such as the requirement to export minerals and other natural resources back to China.

SECTION 15.2 SUMMARY

- Economic development is not the same as economic growth.
- Free-market economists generally argue that reducing barriers to international trade is more important than aid in promoting economic development.
- Strategic trade theorists, by contrast, advocate the case for strategic protection in Global South countries.
- Economic growth has costs as well as benefits, but costs are fewer and benefits greater when successful economic development takes place.
- Most economists agree that some forms of aid are good but that other forms are questionable in terms of promoting economic development.
- Trade may be a more significant cause of development than aid.

Questions

1 Distinguish between a floating and a fixed exchange rate.

2 Is the UK pound's current exchange rate floating or fixed?

3 Would you agree that there has been an exchange rate crisis in the UK in the period before and after Brexit? Justify your answer.

4 Using a demand and supply diagram, explain how a trade deficit is eliminated in a freely floating exchange rate system.

5 Explain the advantages and disadvantages of a floating exchange rate.

6 Distinguish between economic growth and economic development.

7 Explain three factors that might affect the growth and development of a country in sub-Saharan Africa.

8 Do you agree that national income figures provide an inadequate measure of economic development? Justify your answer.

Macroeconomics key terms

absolute advantage a country has an absolute advantage if it can produce more of a good than other countries from the same amount of resources.

accelerator a change in the level of investment in new capital goods induced by a change in national income or output. The size of the accelerator depends on the economy's capital–output ratio.

actual output level of real output produced in the economy in a particular year — not to be confused with the trend level of output, which is what the economy is capable of producing when working at full capacity. Actual output differs from the trend level of output when there are output gaps.

aggregate demand total planned spending on real output in the economy at different price levels.

aggregate supply the level of real national output that producers are prepared to supply at different average price levels.

aid money, goods and services, and 'soft' loans given by the government of one country or a multilateral institution such as the World Bank to help another country. Non-government organisations (NGOs) such as Oxfam also provide aid.

asset a resource (physical or otherwise) that can be sold for money.

automatic stabilisers fiscal policy instruments (e.g. progressive taxes and income-related welfare benefits) that automatically stimulate aggregate demand in an economic downswing and depress aggregate demand in an upswing, thereby 'smoothing' the economic cycle.

availability of credit funds available for households and firms to borrow.

balance of payments a record of all the currency flows into and out of a country in a particular time period.

balance of payments equilibrium a situation in which a deficit or surplus on the current account of the balance of payments is exactly matched by capital inflows or outflows in the other parts of the balance of payments.

balance of primary income comprises inward income flowing into the economy in the current year, generated by UK-owned capital assets located overseas (e.g. shares owned in foreign companies, or land owned overseas), and outward flows of income from the economy in the current year, generated by overseas-owned capital assets located in the UK.

balance of secondary income current transfers (e.g. gifts of money, international aid and transfers between the UK and the EU), flowing into or out of the UK economy in a particular year.

balance of trade the difference between the money value of a country's imports and its exports. Balance of trade is the largest component of a country's balance of payments on current account.

balance of trade deficit when the money value of a country's imports exceeds the money value of its exports.

balance of trade in goods the part of the current account measuring payments for exports and imports of goods. The difference between the total value of exports and the total value of imports is sometimes called the 'balance of visible trade'.

balance of trade in services part of the current account, the difference between the payments for the exports of services and the payments for the imports of services.

balance of trade surplus when the money value of a country's exports exceeds the money value of its imports

balanced budget achieved when government spending equals government revenue ($G = T$).

Bank Rate the rate of interest that the Bank of England pays to commercial banks on their deposits held at the Bank of England (it is also referred to as the base rate).

bonds financial securities sold by companies (corporate bonds) or by governments (government bonds) which are a form of long-term borrowing. Bonds usually have a maturity date on which they are redeemed, with the borrower usually making a fixed interest payment each year until the bond matures).

broad money the part of the stock of money (or money supply) made up of cash, other liquid assets such as bank and building society deposits, but also some illiquid assets. The measure of broad money used by the Bank of England is called M4.

budget deficit occurs when government spending exceeds government revenue ($G > T$).

budget surplus occurs when government spending is less than government revenue ($G < T$).

capital gain the profit made on the buying and selling of financial and other assets (capital losses are the losses made in buying and selling assets).

capital markets where securities such as shares and bonds are issued to raise medium- to long-term financing, and are traded on the 'second-hand' part of the market (e.g. the London Stock Exchange and the New York Stock Exchange).

capital ratio the amount of capital on a bank's balance sheet as a proportion of its loans.

central bank a national bank that provides financial and banking services for its country's government and banking system, as well as implementing the government's monetary policy and issuing currency. The Bank of England is the UK's central bank. The European Central Bank (ECB) implements monetary policy for countries using the euro as their official national currency.

claimant count the method of measuring unemployment according to those people who are claiming unemployment-related benefits (Jobseeker's Allowance).

closed economy an economy with no international trade.

commercial bank (also known as a **retail bank** or **high street bank**) a financial institution which aims to make profits by selling banking services to its customers.

commercial bills short-dated debt (usually three months to maturity), issued by private businesses, which pays the holder a fixed rate of interest.

comparative advantage this is measured in terms of opportunity cost: the country with the least opportunity cost when producing a good possesses a comparative advantage in that good.

consumer prices index the official measure used to calculate the rate of consumer price inflation in the UK. It calculates the average price increase of a basket of 700 different consumer goods and services.

consumption total planned spending by households on consumer goods and services produced within the economy.

contractionary fiscal policy uses fiscal policy to decrease aggregate demand and to shift the *AD* curve to the left.

contractionary monetary policy uses higher interest rates to decrease aggregate demand and to shift the *AD* curve to the left.

corporate bonds debt security issued by a company and sold as new issues to people who lend long term to the company. They can usually be resold second hand on a stock exchange.

corruption a barrier holding back economic growth and development, especially in less developed economies.

cost-push inflation (also known as **cost inflation**) a rising price level caused by an increase in the costs of production, shown by a shift of the *SRAS* (or Keynesian *AS*) curve to the left.

coupon the guaranteed fixed annual interest payment, often divided into two six-month payments, paid by the issuer of a bond to the owner of the bond.

credit when a bank makes a loan, it creates credit. The loan results in the creation of an advance, which is an asset on the bank's balance sheet, and a deposit, which is a liability of the bank.

credit crunch occurs when there is a lack of funds available in the credit market, making it difficult for borrowers to obtain finance and leading to a rise in the cost of borrowing.

crowding out a situation in which an increase in government or public-sector spending displaces private-sector spending, with little or no increase in aggregate demand.

currency union an agreement between a group of countries to share a common currency, and usually to have a single monetary and foreign exchange rate policy.

current account deficit when currency outflows in the current account exceed the currency inflows.

current account measures all the currency flows into and out of a country in a particular time period in payment for exports and imports of goods and services, together with primary and secondary income flows (previously known as 'income flows and transfers').

current account surplus when currency inflows in the current account exceed the currency outflows.

customs unions trading blocs in which member countries enjoy internal free trade in goods and possibly services, with all the member countries protected by a common external tariff barrier.

cyclical budget deficit the part of the budget deficit which rises in the downswing of the economic cycle and falls in the upswing of the cycle.

cyclical unemployment (also known as **Keynesian unemployment** and **demand-deficient unemployment**) unemployment caused by a lack of aggregate demand in the economy, occurring when the economy goes into a recession or depression.

debt people's financial liabilities, or money that they owe.

deficit financing deliberately running a budget deficit and then borrowing to finance the deficit.

deflation a continuing tendency for the average price level to fall.

deindustrialisation the decline of manufacturing industries, together with coal mining.

demand side relates to the impact of changes in aggregate demand on the economy; associated with Keynesian economics.

demand-pull inflation (also known as **demand inflation**) a rising price level caused by an increase in aggregate demand, shown by a shift of the *AD* curve to the right.

demand-side fiscal policy used to increase or decrease the level of aggregate demand (and to shift the *AD* curve right or left) through changes in government spending, taxation and the budget balance.

deregulation involves removing previously imposed regulations. It is the opposite of regulation.

direct tax a tax that cannot be shifted by the person legally liable to pay the tax onto someone else. Direct taxes are levied on income and wealth.

discretionary fiscal policy involves making discrete (separate) changes to *G*, *T* and the budget deficit to manage the level of aggregate demand.

disinflation when the rate of inflation is falling, but still positive.

distribution of income the spread of different incomes among individuals and different income groups in the economy.

economic cycle (also known as a **business cycle** or **trade cycle**) upswing and downswing in aggregate economic activity taking place over 4 to 12 years.

economic shock an unexpected event hitting the economy. Economic shocks can be demand-side or supply-side shocks (and sometimes both) and unfavourable or favourable.

education and training education develops individual knowledge and intellect, while training develops work skills. Both are necessary for economic growth and development.

equilibrium national income the level of income at which withdrawals from the circular flow of income equal injections into the flow; also the level of output at which aggregate demand equals aggregate supply.

equilibrium unemployment exists when the economy's aggregate labour market is in equilibrium. It is the same as the natural *level* of unemployment.

equity the assets that people own.

European Union an economic and partially political union established in 1993. It developed as a more closely integrated organisation from the European Economic Community (EEC), originally formed in 1957. The EU has expanded to include (after the UK's departure) 27 countries.

eurozone (also known as the **euro area**) the name used for the group of EU countries that have replaced their national currencies with the euro. As of 2022, 19 of the 27 EU members have replaced their national currencies with the euro.

exchange rate the external price of a currency, usually measured against another currency.

expansionary fiscal policy uses fiscal policy to increase aggregate demand and to shift the *AD* curve to the right.

expansionary monetary policy uses lower interest rates to increase aggregate demand and to shift the *AD* curve to the right.

expenditure-reducing policy a government policy which aims to eliminate a current account deficit by reducing the demand for imports, by reducing the level of aggregate demand in the economy. Conversely, to reduce a current account surplus, aggregate demand would be increased and spending on imports would rise.

expenditure-switching policy a government policy which aims to reduce a current account deficit by switching domestic demand away from imports to domestically produced goods. Conversely, to reduce a current account surplus, the policy would aim to switch domestic demand away from domestically produced goods towards imports.

export subsidies money given to domestic firms by the government to encourage firms to sell their products abroad and to help make their goods cheaper in export markets.

export-led growth in the short run, economic growth resulting from an increase in exports, which is one of the components of aggregate demand; in the long run, economic growth resulting from the growth and increased international competitiveness of exporting industries.

exports domestically produced goods or services sold to residents of other countries.

financial account the part of the balance of payments which records capital flows into and out of the economy.

Financial Conduct Authority aims to make sure that financial markets work well so that consumers get a fair deal, by ensuring that the financial industry is run with integrity and that consumers can trust that firms have their best interests at heart, and by providing consumers with appropriate financial products and services.

financial markets markets in which financial assets or securities are traded.

Financial Policy Committee the part of the Bank of England charged with the primary objective of identifying, monitoring and taking action to remove or reduce systemic risks with a view to protecting and enhancing the resilience of the UK financial system. The committee's secondary objective is to support the economic policy of the government.

fiscal policy the use by the government of government spending and taxation to try to achieve the government's policy objectives.

fixed exchange rate an exchange rate fixed at a certain value and maintained at this value by the central bank's intervention in the foreign exchange market.

foreign direct investment investment in capital assets (e.g. manufacturing and service industry capacity) in a foreign country by a business with headquarters in another country. Very often the overseas company establishes subsidiary companies in the countries in which it is investing.

foreign exchange markets (also known as forex, FX or currency markets) global, decentralised markets for the trading of currencies.

forward guidance attempts to send signals to financial markets, businesses and individuals, about the Bank of England's interest rate policy in the months and years ahead, so that economic agents are not surprised by a sudden and unexpected change in policy.

free trade area in a free trade area, member countries abolish tariffs on mutual trade, but each partner determines its own tariffs on trade with non-member countries.

freely floating exchange rate the exchange rate is determined solely by the interplay of demand for, and supply of, the currency.

frictional unemployment (also known as **transitional unemployment**) unemployment that is usually short term and occurs when a worker switches between jobs.

full employment according to Beveridge's definition, full employment means 3% or less of the labour force unemployed. According to the free-market definition, it is the level of employment at the market-clearing real-wage rate, where the number of workers whom employers wish to hire equals the number of workers wanting to work.

full employment income the level of income when the economy is producing on its production possibility frontier, with no spare capacity.

globalisation the process of increasing economic integration of the world's economies.

government bonds debt security, in the UK known as gilt-edged securities or gilts, issued by a government and sold as new issues to people who lend long term to the government. They can be resold second hand on a stock exchange.

gross domestic product (GDP) the sum of all goods and services, or level of output, produced in the economy over a period of time, e.g. one year.

hidden economy (also known as the **informal economy**, the **underground economy** and the **black economy**) all the economic transactions conducted in cash which are not recorded in the national income figures because of tax evasion.

human capital the skills, knowledge and experience possessed by the population.

imports goods or services produced in other countries and sold to residents of this country.

index number a number used in an index (e.g. the consumer prices index) to enable accurate comparisons over time to be made.

indexation the automatic adjustment of items such as pensions and welfare benefits to changes in the price level, through the use of a price index.

indicators of development these include gross domestic product (GDP) per head, information on the distribution of income, mortality rates and health statistics.

indirect tax a tax that can be shifted by the person legally liable to pay the tax onto someone else (e.g. through raising the price of a good being sold by the taxpayer). Indirect taxes are levied on spending.

inflation a persistent or continuing rise in the average price level.

infrastructure for the most part, the result of past investment in buildings, roads, bridges, power supplies, fast broadband and other fixed capital goods that are needed for the economy to operate efficiently.

injection spending entering the circular flow of income as a result of investment, government spending and exports.

institutional factors examples include rules, laws, constitutions, the financial system and defined property rights.

interventionist policies occur when the government intervenes in, and sometimes replaces, free markets. Interventionist supply-side policies include government funding of research and development.

investment total planned spending by firms on capital goods produced within the economy.

investment bank a bank which does not generally accept deposits from ordinary members of the general public. Traditional 'investment banking' refers to financial advisory work, such as advising private companies on how to become public companies by floating on the stock market (making shares available for the public to buy or sell), or advising public companies on how to buy up other companies. Investment banks also deal directly in financial markets for their own account.

involuntary unemployment occurs when workers are willing to work at current market wage rates but there are no jobs available.

Keynesian economists followers of the economist John Maynard Keynes, who generally believe that governments should manage the economy, particularly through the use of fiscal policy.

Labour Force Survey a quarterly sample survey of households in the UK. Its purpose is to provide information on the UK labour market. The survey seeks information on respondents' personal circumstances and their labour market status during a period of one to four weeks.

less developed economies countries considered behind in terms of their economy, human capital, infrastructure and industrial base.

liability an amount owed by a business — either a current (short-term) or non-current (long-term) liability

life-cycle theory of consumption a theory that explains consumption and saving in terms of how people expect their incomes to change over the whole of their life cycles.

liquidity measures the ease with which an asset can be converted into cash without loss of value. Cash is the most liquid of all assets.

liquidity ratio the ratio of a bank's cash and other liquid assets to its deposits.

long-run aggregate supply the real output that can be supplied when the economy is on its production possibility frontier. This is when all the available factors of production are employed and producing at their 'normal capacity' level of output.

long-run economic growth an increase in the economy's potential level of real output, and an outward movement of the economy's production possibility frontier.

macroeconomic indicators provide information from recent economic performance for judging the success or failure of a particular type of government policy (e.g. fiscal policy or monetary policy).

managed exchange rates where the government intervenes in the foreign exchange market to influence its value.

marginal propensity to consume the fraction of any increase in income which people plan to spend on the consumption of domestically produced goods and services.

marginal propensity to save the fraction of any increase in income which people plan to save rather than spend.

market-based supply-side policies (or **non-interventionist supply-side policies**) these policies free up markets, promote competition and greater efficiency, and reduce the economic role of the state.

marketisation (also known as **commercialisation**) involves shifting provision of goods or services from the non-market sector to the market sector.

maturity date the date on which the issuer of a dated security, such as a gilt-edged security (long-dated) or a Treasury bill (short-dated), pays the face value of the security to the security's owner.

monetarism narrow monetarism centres on increases in the money supply as the prime cause of inflation; broader monetarism focuses on the virtues of free markets in resource allocation.

monetarists economists who argue that a prior increase in the money supply is the cause of inflation.

monetary policy the use by the government and its agent, the Bank of England, of interest rates and other monetary instruments to try to achieve the government's policy objectives.

Monetary Policy Committee a group of Bank of England officials and independent experts who meet to set interest rates in the UK eight times a year.

monetary policy instruments tools such as Bank Rate which are used to try to achieve monetary policy objectives.

money primarily a medium of exchange or means of payment, but also a store of value.

money markets provide a means for lenders and borrowers to satisfy their short-term financial needs. The term 'money market' covers several markets, including those for Treasury bills and commercial bills.

money supply the stock of financial assets which function as money.

moral hazard the tendency of individuals and firms, once protected against some contingency, to behave so as to make that contingency more likely.

more developed economies countries with a high degree of economic development, high average income per head, high standards of living, usually service industries dominating manufacturing, and investment having taken place over many years in human capital and infrastructure.

multinational corporations enterprises operating in several countries but with their headquarters in one country.

multiplier the relationship between a change in aggregate demand and the resulting generally larger change in national income.

narrow money the part of the stock of money (or money supply) made up of cash and liquid bank and building society deposits.

national capital stock the stock of capital goods (e.g. buildings and machinery) in the economy that has accumulated over time and is measured at a point in time.

national debt the *stock* of all past government borrowing that has not been paid back.

national income the *flow* of new output produced by the economy in a particular period (e.g. a year), measured by the flow of factor incomes.

national product (also known as **national output**) the *flow* of new output produced by different industries in a particular period (e.g. a year).

national wealth the stock of all goods that exist at a point in time that have value in the economy.

natural rate of unemployment the rate of unemployment when the aggregate labour market is in equilibrium.

negative output gap the level of *actual* real output in the economy is lower than the *trend* output level.

nominal GDP GDP measured at the current market prices, without removing the effects of inflation.

normal capacity level of output the level of output at which the full production potential of the economy is being used.

Office for Budget Responsibility advisory public body that provides independent economic forecasts and analysis of the public finances as background to the preparation of the UK budget.

open economy an economy open to international trade.

output gaps show the level of *actual* real output in the economy either higher or lower than the *trend* output level.

Phillips curve based on evidence from the economy, this shows the apparent relationship between the rate of inflation and the rate of unemployment; now known as 'the short-run Phillips curve'.

policy conflict when two policy objectives cannot both be achieved at the same time: the better the performance in achieving one objective, the worse the performance in achieving the other.

policy objective a target or goal that policy-makers aim to 'hit'.

portfolio investment the purchase of one country's securities (e.g. bonds and shares) by the residents or financial institutions of another country.

positive output gap the level of *actual* real output in the economy is greater than the *trend* output level.

price index an index number showing the extent to which a price, or a 'basket' of prices, has changed over a month, quarter or year, in comparison with the price(s) in a base year.

principle of taxation (also known as a **canon of taxation**) a criterion used for judging whether a tax is good or bad.

privatisation involves shifting ownership of state-owned assets to the private sector.

profitability the state or condition of yielding a financial profit or gain.

pro-free-market economists opponents of Keynesian economists, who dislike government intervention in the economy and who much prefer the operation of free markets.

proportional taxation when the proportion of income paid in tax stays the same as income increases.

Prudential Regulation Authority the part of the Bank of England responsible for the microprudential regulation and supervision of banks, building societies, insurers and major investment firms.

purchasing power parity (PPP) exchange rates the rates of currency conversion that equalise the purchasing power of different currencies by eliminating the differences in price levels between countries.

quantitative easing when the Bank of England buys assets, usually government bonds, with money that the Bank has created electronically.

quantity theory of money the oldest theory of inflation, which states that inflation is caused by a persistent increase in the supply of money.

quotas physical limits on the quantities of imported goods allowed into a country.

rate of interest the reward for lending savings to somebody else (e.g. a bank) and the cost of borrowing.

real GDP a measure of all the goods and services produced in an economy, adjusted for price changes or inflation. The adjustment transforms changes in nominal GDP, which is measured in money terms, into a measure that reflects changes in the total output of the economy.

real wages the purchasing power of the nominal (or money) wage; for example, real wages fall when inflation is higher than the rise in the nominal wage rate, and real wages rise when the nominal wage rate increases more rapidly than inflation.

real-wage unemployment unemployment caused by real wages being stuck above the equilibrium market-clearing real wage.

recession in the UK and many other countries, a recession is defined as six months or more of negative economic growth or declining real national output.

reflationary policies policies that increase aggregate demand with the intention of increasing real output and employment.

regressive taxation when the proportion of income paid in tax falls as income increases.

retail prices index a measure formerly used to calculate the rate of consumer price inflation in the UK.

saving income which is not spent.

seasonal unemployment unemployment arising in different seasons of the year, caused by factors such as the weather.

security secured loans, such as mortgage loans secured against the value of property, are less risky for banks than unsecured loans.

shares undated financial assets, sold initially by a company to raise financial capital. Shares sold by public limited companies (PLCs) are marketable on a stock exchange, but shares sold by private limited companies (Ltds) are not marketable. Unlike a loan, a share signifies that the holder owns part of the enterprise.

short-run economic growth growth of real output resulting from using idle resources, including labour, thereby taking up the slack in the economy.

short-term speculative capital (also known as **hot money**) refers to flows of money across international borders, moving to find the greatest short-term return; it is usually very easy to move and will move quickly as returns and opportunities arise.

structural budget deficit the part of the budget deficit which is not affected by the economic cycle but results from structural change in the economy affecting the government's finances, and also from long-term government policy decisions.

structural unemployment long-term unemployment occurring when some industries are declining, even though other industries may be growing. It also occurs within a growing industry if automation reduces the demand for labour, and when production requires new skills not possessed by the workers who lose their jobs. Structural unemployment is associated with the occupational and geographical immobility of labour.

supply side relates to changes in the potential output of the economy, which is affected by the available factors of production (e.g. changes in the size of the labour force and the productivity of labour).

supply-side economics a branch of free-market economics arguing that government policy should be used to improve the competitiveness and efficiency of markets and, through this, the performance of the economy.

supply-side fiscal policy used to increase the economy's ability to produce and supply goods, through creating incentives to work, save, invest and be entrepreneurial. Interventionist supply-side fiscal policies, such as the financing of retraining schemes for unemployed workers, are also designed to improve supply-side performance.

supply-side improvements reforms undertaken by the private sector to increase productivity so as to reduce costs and to become more efficient and competitive. Supply-side improvement often results from more investment and innovation, often undertaken by firms without prompting from the government.

supply-side policies government economic policies which aim to make markets more competitive and efficient, increase production potential and shift the *LRAS* curve to the right. Supply-side fiscal policy is arguably the most important type of supply-side policy, but there are also non-fiscal supply-side policies.

systemic risk in a financial context, the risk of a breakdown of the entire financial system, caused by inter-linkages in the financial system, rather than simply the failure of an individual bank or financial institution.

tariffs (also known as **import duties**) taxes imposed on imports from other countries entering a country.

total managed expenditure the total amount that the government spends. It splits into the amount that government departments (e.g. Defence) have been allocated to spend and spending that is not controlled by a government department, including welfare, pensions and national debt interest payments.

trade-off between policy objectives although it may be impossible to achieve two desirable objectives at the same time (e.g. zero inflation and full employment), policy-makers may be able to choose an acceptable combination lying between the extremes, e.g. 2% inflation and 4% unemployment.

Treasury bills short-dated government debt (usually three months to maturity), which pays the holder a fixed rate of interest until redemption.

trend growth rate the rate at which output can grow, on a sustained basis, without putting upward or downward pressure on inflation. It reflects the annual average percentage increase in the productive capacity of the economy.

United Nations Human Development Index an index based on life expectancy, education and per capita income indicators, which ranks the world's countries into four tiers of human development: (i) very high human development; (ii) high human development; (iii) medium human development; (iv) low human development.

voluntary unemployment occurs when workers choose to remain unemployed and refuse job offers at current market wage rates.

withdrawal a leakage of spending power out of the circular flow of income into savings, taxation or imports.

World Trade Organization an international body whose purpose is to promote free trade by persuading countries to abolish import tariffs and other barriers to trade. As such, it has become closely associated with globalisation.

yield the annual interest on a bond expressed as a percentage of the bond's current market price.

Practice questions

See www.hoddereducation.co.uk/aqa-economics-alevel for guidance on answering the practice questions.

Paper 1: Markets and market failure

Context 1

Total for this context: 40 marks

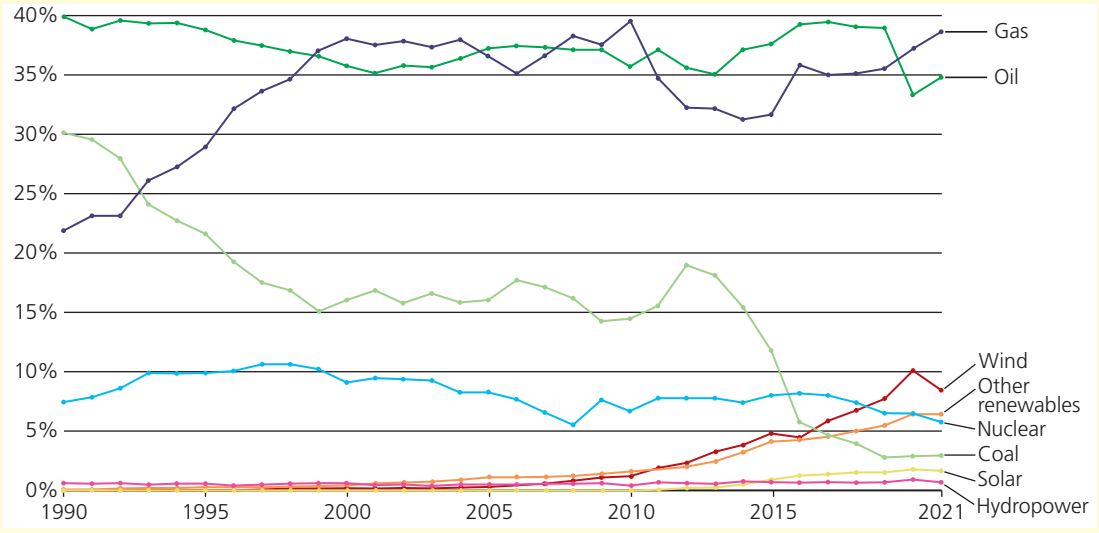
The UK energy market

Study Extracts **A**, **B**, **C** and **D**, and then answer **all** parts of Context 1 which follow.

Extract A: Number of major power producers in the UK

Year	Number	Number producing at least 5% of total generation
1989	6	–
1991	11	–
1995	25	–
2001	36	6
2005	30	7
2010	39	8
2015	53	6
2020	58	6

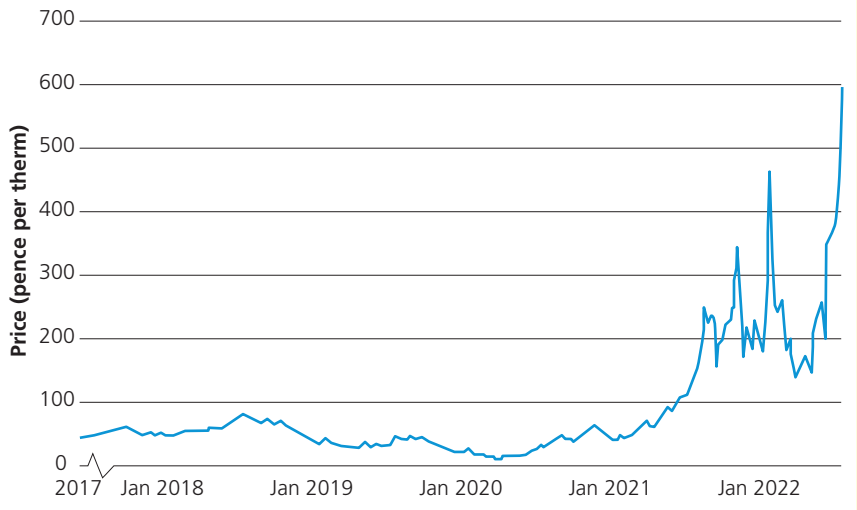
Extract B



Note: to convert from primary direct energy consumption, an inefficiency factor has been applied for fossil fuels (i.e. the 'substitution method').

Source: Our World in Data based on *BP Statistical Review of World Energy* (2022)

Figure 1 Share of energy consumption by source, UK, 1989–2021



Source: BBC market data

Figure 2 Natural gas price, 2017–22

Extract C: Structure of the UK's energy markets

In the 1980s, UK government policy sought to liberalise the energy market by breaking up monopoly power and creating competition. In the decades before the 1980s, energy companies were state-owned natural monopolies. Privatisation and deregulation allowed new entrants into the market. Energy firms were no longer monopolies but had to compete for retail customers. Market economists welcomed the changes to market structure, arguing that it would lead to lower prices and greater efficiency. In the short term, government regulators would intervene to promote competition, but in the long term, it was hoped the need for government would disappear.

Prices did fall in the 2000s but the market was dominated by a few large firms, known as the 'Big Six' (British Gas, EDF Energy, E.ON, RWE npower, Scottish Power and SSE). High levels of boardroom pay, excessive profits and generous dividend payments led to public anger. The government's response was to seek further liberalisation to make the industry more competitive. The market was opened up to allow new, smaller firms that did not produce energy but simply bought it on the global wholesale markets and sold it to retail customers. When wholesale prices were cheap, they offered consumers choice and lower prices. However, when wholesale prices began to rise rapidly, these thinly capitalised, smaller firms quickly went out of business.

Critics of the energy market have argued that privatised energy firms seek short-term profits and lack the incentive to invest in the long term. In the good years, increased competition allowed consumers to benefit from cheap energy, but insufficient expensive investment projects were commissioned. The UK's ageing nuclear sector has been replaced, there is very little capacity to store gas, and battery storage capacity is still in its infancy. The UK is highly dependent on imported gas in the winter months. Some critics believe the government should embark on a programme of nationalisation and government planning to protect the public interest. Free-market economists reject this approach. They concede that the system has its flaws but point to the investment undertaken by private companies and the UK's record of successful decarbonisation.

Source: News reports

Extract D: Decarbonising the supply of power

The decarbonisation of the UK's supply of energy started in the 1980s with the decline of the coal industry and the privatisation of the energy market. The newly privatised power companies wanted to move away from the relatively expensive coal and switch to cheaper gas and oil alternatives. Burning gas produces half as many CO_2 emissions as coal, which has helped the UK significantly reduce its greenhouse gas emissions. In 2013, the government introduced a carbon tax, which hit coal twice as hard as gas. Today, burning coal accounts for less than 2% of the UK's energy supply. By eliminating coal from its energy mix, the UK has reduced its carbon emissions faster than any other developed country.

The UK has been keen to invest in renewables, such as wind and solar. The wind turbines that have been built both onshore and offshore in the last decade now supply over 25% of the UK's energy on a good day. There are plans to keep building, but there are two major problems with this strategy. First, wind turbines are often unpopular in the local communities in which they are built and struggle to gain planning permission. Secondly, wind power is intermittent and on still days the national grid is dependent on gas-burning power stations to keep the lights on.

Nuclear power is another viable source of reliable, low-carbon energy. However, successive governments have been unwilling to finance the building of expensive new nuclear power plants. Furthermore, the public have been nervous of a sector which has experienced major disasters, such as Fukushima in Japan in 2011.

The UK is still reliant on gas, which accounts for 40% of its electricity supply and heats 85% of its homes. Until recently, gas was seen to be a viable bridging fossil fuel while new greener technologies were being developed. Recent rapid price surges in gas have brought the role of the fuel into doubt. Increased global demand and unstable gas supply from autocratic regimes have left the UK facing an energy crisis.

Source: News reports

[01] Using the data in **Extract A**, calculate the percentage change in the number of major power producers in the UK between 1989 and 2020. Give your answer to **1** decimal place. *(2 marks)*

[02] Explain how the data in **Extract B** show UK households are vulnerable to a price shock in the energy market. *(4 marks)*

[03] **Extract C** states 'In the decades before the 1980s, energy companies were state-owned natural monopolies. Privatisation and deregulation allowed new entrants into the market. Energy firms were no longer monopolies but had to compete for retail customers.'
Using a diagram, explain how increased competition can lead to lower prices for consumers **and** greater efficiency. *(9 marks)*

[04] **Extract D** states 'Recent rapid price surges in gas have brought the role of the fuel into doubt. Increased global demand and unstable gas supply from autocratic regimes have left the UK facing an energy crisis.'
Evaluate the view that the best way to protect the public interest in the long run is for the government to nationalise the energy industry. *(25 marks)*

Essay 1

Since 1980, inequality in the UK has increased but successive governments have been committed to cut income tax in the long run.

(a) Explain how cutting income tax will affect equality and equity. *(15 marks)*

(b) Evaluate the view that market-based policies are the best way to raise the incomes of all households in the UK. *(25 marks)*

Essay 2

For a company to grow it must reinvest into the business. Aggressive investment and expansion strategies have seen Amazon become one of the largest companies in the world. In 2021, Amazon posted revenues of over $469 billion but posted profits of just over $33 billion.

(a) Explain how a firm can both lower average costs **and** increase sales by investing in its business. *(15 marks)*

(b) Evaluate the view that, in the modern economy, the assumption that the main objective of the firm is to maximise profits no longer holds. *(25 marks)*

Essay 3

Earth's temperature has risen by 0.14°F (0.08°C) per decade since 1880, but the rate of warming since 1981 is more than twice that, at 0.32°F (0.18°C) per decade. 2021 was the sixth warmest year on record.

(a) Explain how the lack of property rights can lead to the destruction of common resources. *(15 marks)*

(b) Assess the view that taxation is the most effective policy tool available to governments in reducing greenhouse gas emissions. *(25 marks)*

Paper 2: National and international economy

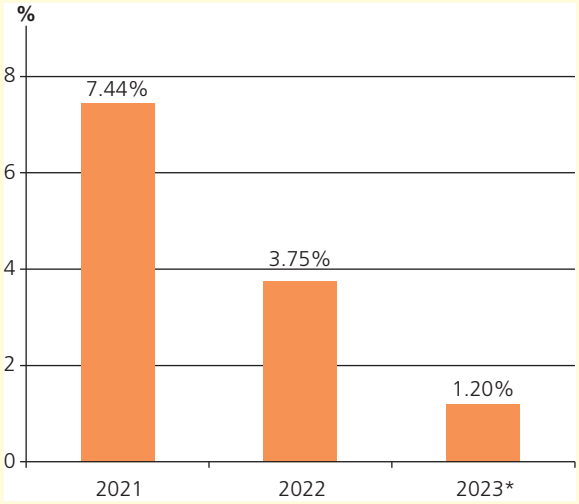
Context 2

Total for this context: 40 marks

The return of stagflation

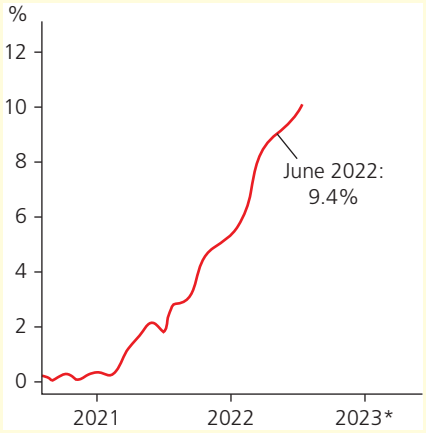
Study Extracts **A**, **B** and **C**, and then answer **all** parts of Context 2 which follow.

Extract A



* forecast

Figure 1 Annual change in real GDP, UK, 2021–23



* forecast

Figure 2 CPI annual inflation rate, UK, 2021–23

Extract B: Fine-tuning the economy

Once upon a time, macroeconomic policy seemed very easy to implement. All the policy-makers had to do was to fine-tune the level of aggregate demand in the economy. When economic growth was too sluggish or when unemployment was rising, the Treasury could cut taxes (expansionary fiscal policy) or the Bank of England could cut interest rates (expansionary monetary policy). Both of these would inject demand into economy, which via the multiplier process would deliver economic growth and full employment.

If eventually the economy overheated with rising prices and labour shortages, the policy could simply be thrown into reverse. Contractionary fiscal and monetary policies could take demand out of the economy until normal service was resumed. It was all a matter of timing. If the policy-makers got their timing right and were not too greedy when setting their targets, the economy could move upwards and upwards without suffering any major economic problem.

Policy-makers were saying to governments 'allow us to manage demand and fine-tune the economy, and we will deliver economic growth, more jobs and low inflation'. But in the 1970s, a new problem emerged which became known as 'stagflation'. The two economic evils of falling real GDP and accelerating inflation — which, with 'smart' economic management of the economy, were not supposed to occur together — did indeed occur together.

Source: News reports

Extract C: Analysing stagflation

In recent years, the fear of stagflation has re-emerged, as have newspaper articles worrying about the problem. For example, on 1 August 2022, the *London Evening Standard* published an article titled 'Inflation and recession: we're taking a trip back to the chaotic 1970s'.

Eminent economists are also trying to explain and analyse stagflation and to suggest policies that could be used to reduce and eventually to eliminate it. Robert Carnell, the Regional Head of Research at Asia-Pacific, has identified various stages in the stagflationary process which are illustrated in Figure 3.

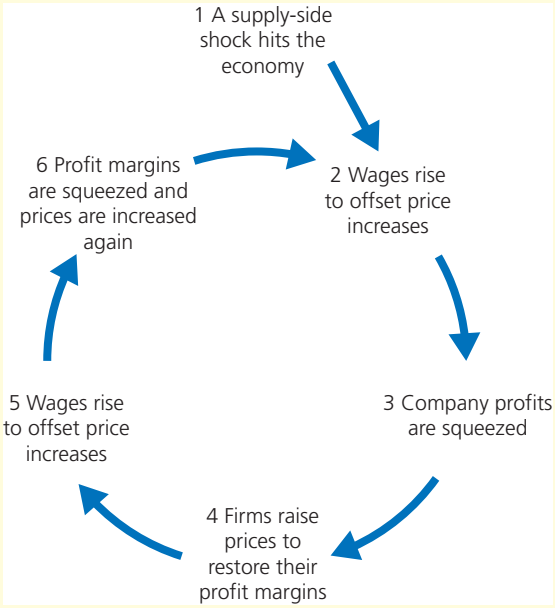


Figure 3 The stagflationary process

In Figure 3, stagflation kicks off with an unexpected supply-side shock hitting the economy. Recent examples of such supply-side shocks include the Covid-19 health pandemic and the Russian invasion of Ukraine. The initial supply-side shock, shown by Item 1 in Figure 3, triggers a process of cost-push inflation which is illustrated in the later items in the diagram.

Source: News reports

[01] In 2020, nominal UK GDP was estimated to be £2.04 trillion. Using the data in **Extract A (Figure 1)**, calculate the nominal UK GDP for 2021. *(2 marks)*

[02] Explain how the data in **Extract A** show that the UK economy is going to experience the economic problem of stagflation. *(4 marks)*

[03] **Extract C** states that 'Recent examples of such supply-side shocks include the Covid-19 health pandemic and the Russian invasion of Ukraine.'
With the help of a diagram, explain how a supply-side shock affects the rate of inflation. *(9 marks)*

[04] **Extract B** states that 'The two economic evils of falling real GDP and accelerating inflation — which, with 'smart' economic management of the economy, were not supposed to occur together — did indeed occur together.'
Evaluate the view that an aggressive contractionary monetary policy is the best way to eliminate stagflation in the UK economy today.
(25 marks)

Essay 1

'Cutting taxes is the best chance the country has of generating growth. It is the only way to secure Britain's economic recovery.' (Professor Patrick Minford)

(a) Explain how large tax cuts will affect the budget deficit and the national debt. *(15 marks)*
(b) Assess the view that, as long as the economy grows, large budget deficits are not a problem. *(25 marks)*

Essay 2

More and more UK firms are looking into reshoring production (bringing manufacturing operations back to the UK) following supply chain problems and uncertainty caused by the Covid-19 pandemic and Brexit.

(a) Explain how UK firms' reshoring production will affect the UK's trade balance in goods. *(15 marks)*
(b) Evaluate the view that shorter supply chains and less international trade will enable the UK to achieve its macroeconomic objectives. *(25 marks)*

Essay 3

Since 2015, the pound sterling has lost 20% of its value against the US dollar and 15% against the euro.

(a) Explain three policies which can influence the value of a currency in a floating exchange rate system. *(15 marks)*
(b) Evaluate the view that the significant depreciation of the currency against its major trading partners will improve the UK's economic performance. *(25 marks)*

A-level Paper 3 investigation question

Source booklet

Poverty in the UK

Extract A: Absolute and relative poverty

There are different approaches to defining and measuring poverty. Two commonly used measures are absolute and relative poverty.

Absolute poverty is when household income is below a certain level. This makes it impossible for the person or family to meet basic needs of life including food, shelter, safe drinking water, education, healthcare, etc.

Relative poverty is when households receive 50% less than average household incomes. So they do have some money but still not enough money to afford anything above the basics. Relative poverty is sometimes described as 'relative deprivation' because the people falling into this category are not living in absolute poverty. They are not, however, enjoying the same standard of life as everyone else in the country. They may be missing out on TV, internet, clean clothes, a safe home (a healthy environment, free from abuse or neglect), or even education.

On the whole, poverty is about exclusion. In its most extreme form, it is the inability to access what you need for a decent life. Relatively speaking, in more developed countries, it is being excluded from what constitutes normal daily life, such as

- internet to access jobs or public services
- the proper clothes to find that job
- access to decent housing.

Before the Covid-19 pandemic, more than one in five people in the UK lived in poverty according to the Joseph Rowntree Foundation. In 2018, the United Nations Special Rapporteur on extreme poverty and human rights expressed great concern that in the UK '14 million people, a fifth of the population, live in poverty. Four million of these are more than 50% below the poverty line, and 1.5 million are destitute, unable to afford basic essentials.'

The cost of living crisis is hitting people on low incomes hardest, putting immense pressure on already stretched household finances. The Food Foundation has found that an increasing number of people in the UK are facing situations in which they cannot afford to meet their basic needs.

The Institute for Fiscal Studies has found that inflation for those on low incomes was three percentage points higher than that being experienced by wealthier families. People are going without food to save money for rent, disconnecting gas to save money, not cooking food to save energy costs. People also report an increasing sense of pressure from trying to support family members and a lot of fear for the future.

The Covid-19 pandemic has had a significant and unequal impact on people's finances — with a sharp increase in claims for Universal Credit in April 2020, and a record 2.5 million emergency food parcels given to people in crisis by the Trussell Trust between April 2020 and March 2021. Almost 700,000 people had been driven into poverty as a result of the pandemic by winter 2020, including 120,000 children.

The predicted long-term effects of the pandemic include high unemployment that pushes many families into poverty, at a scale 10 times greater than the 2008 financial crisis.

Child poverty is defined by the Child Poverty Action Group as existing when parents cannot afford to pay for basics such as food, housing and clothes — it affects one in four children in the UK. A total of 46% of children from black and minority ethnic groups are in poverty in the UK, compared with 26% of children in white British families.

Living without the security of decent housing with constant pressure created by low-paid work or unemployment has significant consequences for emotional and psychological well-being, health, family relationships and the lives of children.

Research from the Joseph Rowntree Foundation (2015) has shown that the experience of poverty significantly affects the way we think, feel and act. It can crush individual confidence, hope, and ambition. It stifles people's ability to participate in our society.

Source: based on information from Habitat for Humanity

Extract B: The main causes of world poverty

The charity Tiyara (www.tiyara.org) lists the seven most common causes of poverty:

1 Lack of access to food and clean water. Without such access, it is hard to have the energy to work. Lack of nutrition also makes people more vulnerable to preventable diseases.

2 Warfare/conflict. War and conflict have a negative effect on economic growth and financial markets. They lower investor confidence, which can lead to job losses. Conflicts may also create population displacement, which can result in severe hardship, with a lack of basic necessities for survival.

3 Lack of jobs. Another cause of poverty is a lack of jobs in a certain area. This can be a result of various factors, including conflict, geographic location and recessions, with rural areas often providing fewer employment opportunities. When there are not enough jobs available, people are unable to make a living and break the cycle of poverty.

4 Social injustice. Another cause of poverty is social injustice, when specific groups are discriminated against on the basis of their race, nationality, caste, gender, sexual orientation or religious beliefs. Such discrimination may prevent individuals in these groups from gaining a good education, which could then affect employability and access to resources that can enable them to break the cycle of poverty.

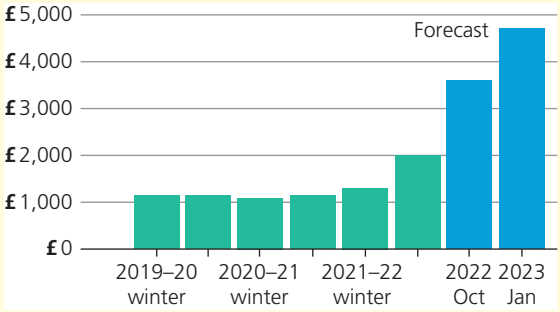
5 Lack of infrastructure. Infrastructure includes roads, bridges, access to public transport and internet access. These are all important in allowing people to travel and communicate. Poor infrastructure can result in people being unable to access jobs. Some people are forced to spend large amounts of time accessing basics such as food and water, because of poor infrastructure.

6 Climate change. Storms, floods and droughts resulting from climate change can have a devastating effect on communities. Agricultural areas are particularly hard hit, because they may lose their source of income as well as their food supply. Such weather catastrophes can create major difficulties even in successful economies.

7 Lack of education. One of the main causes of poverty is the lack of a good education. A good education can enable a person to obtain a job, negotiate a good salary and move themselves out of poverty. Access to a good education is not universal and there are many areas around the world where children do not receive the opportunities they need to do well in life. 30

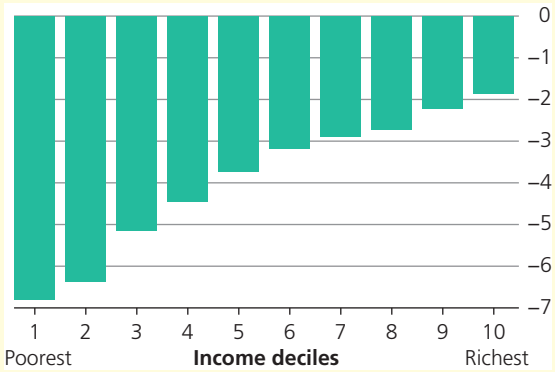
Source: Tiyara

Extract C: Selected economic data



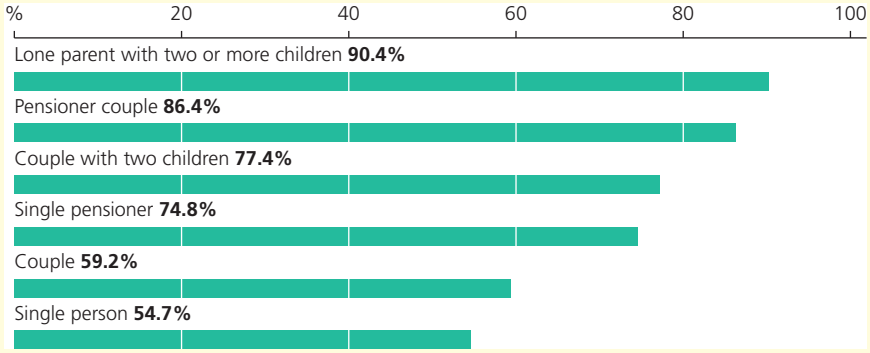
Source: Ofgem

Figure 1 Annual energy bill for a typical UK household, 2019–23



Source: ONS

Figure 2 Fall in money available for discretionary spending by income decile as a result of energy price rises (%)



Source: ONS

Figure 3 Percentage of each family type falling into fuel poverty in January 2023

Extract D: Fuel poverty in the UK

Fuel poverty relates to households that have to spend a high proportion of their household income to keep their home at a reasonable temperature. Fuel poverty is affected by three key factors: a household's income, their fuel costs and their energy consumption (which in turn can be affected by the energy efficiency of the household's dwelling). Two-thirds of households in the UK, or 45 million people, will have been pushed into fuel poverty by January 2023, according to a new report authored by York academics. The report predicts that, even with the £400 fuel rebate being offered by the government, 70.6% of households will be plunged into fuel poverty in Yorkshire and the Humber, 56.4% in London and 76.3% in Northern Ireland.

People living in the poorest and coldest regions of the UK will be the worst affected along with those who are already most likely to be struggling with the cost of living. More than 80% of large families, lone parents and pensioner couples will be in fuel poverty, the authors of the report calculate.

Fuel poverty is defined as having to spend over 10% of net income on fuel. The report estimates that households living in fuel poverty will be spending on average £37.51 above this 10% threshold per month.

Co-lead author of the report Professor Jonathan Bradshaw from the Social Policy research unit at the University of York said: 'This should be the main preoccupation of the Conservative leadership candidates. A new package of mitigations is urgently needed.'

The electricity and gas price cap will rise again in October 2022 and again in January 2023. The size of the increase has not yet been announced but it is expected to take average electricity and gas bills to £64.59 per week or £3,359 per year. It is also expected that the January increase will take weekly bills to £4,200 per year, with some sources predicting even larger increases.

The report highlights how the £400 rebate being offered by the government will reduce these bills by only £15.38 per week between October 2022 and April 2023.

Source: News reports

Section B

Answer all questions in this section.

Refer to source booklet for **Extracts A**, **B**, **C** and **D**. Total for this investigation: 50 marks

Fuel poverty

INVESTIGATION

Scenario

You are an economist working for a UK charity, the Poverty Action Group, which is worried about increased levels of poverty following the Covid-19 health crisis in the UK. The charity is particularly concerned by the growth in fuel poverty or energy poverty triggered by rising energy costs brought about by Russia's invasion of Ukraine in 2022. The charity has requested that you provide answers to three key questions.

Referring to the source booklet, study **Extracts A**, **B** and **C**, and then use these and your own economic knowledge to help you answer **questions 01 and 02**. There is also an additional news report **Extract D** which is to be used with the other Extracts to help you answer **question 03**.

[01] To what extent do the data support the view that the poorest groups in the UK are struggling to buy anything other than basic goods? You must use the data in **Extract C** to support your assessment. *(10 marks)*

[02] Explain the main causes of poverty in the UK. *(15 marks)*

[03] Taking into account the news report concerning fuel poverty in **Extract D**, and the other evidence in **Extracts A**, **B** and **C**, recommend **two** policies the UK government should adopt to reduce or eliminate fuel or energy poverty. Justify your recommendations. *(25 marks)*

Note: page numbers in **bold** refer to key term definitions. Abbreviations used: LDEs for less developed economies; MDEs for more developed economies.

A

abnormal profit **109**
  in monopoly (monopoly profit) 136, 159
  in perfect competition 131–32, 155
above-normal profit **109**
absolute advantage **521–23**
  Smith's and Ricardo's ideas 526
  vs competitive advantage 523–25
absolute poverty **210**
accelerator **350**
  interaction with multiplier 378
  numerical example 351
actual output 376
  and output gaps 380
AD/AS analysis 336
  current account 554–55
  output gaps 380
  and rate of inflation 364
adaptive expectations theory 405, 416
adjustable peg exchange rates 573, 583
administrative costs 285–86, 523
adverse selection 261
advertising
  of demerit goods 259–60
  informative and persuasive 143–44
Africa
  climate change impact 342
  economic development 603–04
  problem of capital flight 601
aggregate demand **336** *see also* inflation
  and cyclical unemployment 387–88
  deficient, effect of 334, 382, 554
  determinants of 343–52
  and economic activity 353
  equation 336–37, 367, 444, 467–68
  and expansionary monetary policy 446
  and expenditure-reducing policy 558
  fiscal policy influencing 464–69
  higher interest rates decreasing 446
  and the level of economic activity 352–58
  meaning of 336–37
  multiplier effect 353–57
aggregate supply **336**
  fiscal policy influencing 470
  meaning of 338
agricultural markets
  fluctuating prices 72
  labour shortages 170–71
aid **601**
  effective and ineffective 603
  forms of 602–03
  vs trade 603–04
air pollution, 1952 great smog 243–44
Akerlof, George, market for 'lemons' 32
allocative efficiency **133**, 134–35
  applied to rail fares 251
  and negative externalities 250–51
  and perfect competition 133
  and static efficiency 159
allocative function of prices **226**
allocative inefficiency **135**
  disadvantage of monopoly 138–39
altruism **39**
anchoring, cognitive bias **37**
Apple
  competition 156
  and creative destruction 117
  'first mover' advantage 122
  and innovation 113
  product differentiation 123–24
artificial barriers **122–23**
  deregulatory policies to remove 268, 276
  and movement between jobs 181
assets **418–19** *see also* money; wealth
  of banks 435–38, 439, 458–60
  capital gains and losses 429
  private pension 203–04
  and quantitative easing 449–50
  risky, sub-prime mortgages 461
  security of 440
  'toxic' 551–2
asymmetric information **31**
  in imperfectly competitive markets 262–64
  in the second-hand car market 32, 262
auctions 68
automatic enrolment 221
automatic stabilisers **490–91**
availability bias **37**
availability of credit **346**
average cost **91**
average cost of labour **176**
average fixed cost (AFC) **91**, 92–93
average returns of labour **87**
  relationship with marginal returns 87, 88
average revenue (AR) **103**
  and firm's demand curve 103
  relationship with marginal revenue 104–07
  short-run profit maximisation in perfect competition 131
average total cost (*ATC*) **91**, 92–93
average variable cost (*AVC*) **91**, 92–93
aviation industry
  jumbo jets, duopoly 116
  reduction of price controls 157

B

bad/malign deflation 407
'bad' taxes 481
'bads' and 'goods' *see* 'goods' and 'bads'
balance of payments **306**, **541** *see also* current account
  equilibrium, floating exchange rate **310**, 576
  financial account: capital flows between countries 550–53
  policy objective 306–07
balance of primary income **543**
balance of secondary income **543**
balance of trade **306**
balance of trade deficit **306**
balance of trade in goods **545–46**
balance of trade in services **547**
balance of trade surplus **306**
balanced budget **464**
  Keynes's views on 466
Bank of England 442–43 *see also* monetary policy; quantitative easing
  banknote liability of 418–19
  and cost-push inflation 400
  credibility of 404
  exchange rate intervention 583, 584
  and financial regulation 456–57, 458
  and fixed exchange rate 580–81, 582
  independence of 397
  lender of last resort 438, 443
  macroprudential policy 457
  and moral hazard 458–59
Bank Rate **444**
  'conventional' and 'unconventional' monetary policy 448
  MPC considerations when setting 445
  and transmission mechanism of interest rate policy 447
  using to manage demand for credit 444
banks 432–34 *see also* central banks
  creation of credit and deposits 435–38
  failures 458–59
  liquidity and capital ratios 459–60
  systemic failure 460–61
barriers to growth and development 597–600
  corruption 597
  institutional factors 598
  lack of human capital 599
  lack of property rights 599–600
  poor infrastructure 598–99
barriers to market entry and exit **121–23**
  absence of in competitive markets 46, 104
  deregulatory policies to remove 157, 268, 276
barter 83, 420, 422
behavioural economics **33**
  altruism and fairness 39
  and automatic enrolment 221
  bounded rationality and self-control 35–36
  and charitable giving 43
  choices and framing 40–42
  decision making biases 35, 37–39

and economic policy 40–43
encouraging energy-saving behaviour 242
fast and slow thinking 36
nudge theory 38, 42–43
and traditional economic theory 34
Beveridge, definition of full employment 301–02, 389
biases in decision-making 35, 37–39
Big Mac index, *The Economist* 326
biofuel production and food prices 71
black (hidden) economy 221, 283, **323**
price controls creating 286
bonds **424**
and capital gains and losses 429
corporate and government 427
portfolio investment by UK residents 551
prices of, inverse link to interest rates 428–29
role in QE 449–51
selling of corporate 430–31
borrowing
government 487, 488, 489, 502–03
household 345, 446
inter-bank 441
sub-prime 348, 460–61
bounded rationality **35–36**
bounded self-control **36**
Brexit referendum (2016)
and devaluation of the pound 584, 585
effect on inflation 396
and 'QE4' 452
broad money **423**
budget day 467
budget deficit **464**
benefits and costs of 493
cyclical and structural 489
and the national debt 487–89
budget history 467
budget surplus **464**
business cycle *see* economic cycle

C

call centres 201, 385, 515
canon of taxation **481**
capital account, balance of payments 542
capital flight 601
capital flows, financial account 543, 550
China to USA 568
and exchange rates 578
long-term direct 551
long-term portfolio 551–52
short-term speculative 552
capital gains 203, **429**
tax levied on 478
capital goods **11**
firms' investment in new 349–50, 446
and national capital stock 321
vs consumer goods, *PPF* diagrams 17, 21
capital markets **426**, 427, 430–31
and portfolio investment 551
primary and the secondary 430–31
supply-side policies 507–08
vs money markets 427
capital mobility 515–16
capital–output ratio, accelerator 350–51
capital productivity **78**, 93
capital ratio **459–60**
capital stock 321–22
investment in 349, 350
car industry *see also* roads/road use
and asymmetric information 32
economies of scale less important 99
and global production platforms 518–19
lean manufacturing 81
new and second-hand market 73
wage discrimination 195
Carney, Mark, forward guidance 453–54
cartels **147**
joint-profit maximisation 148–49
and restrictive trading practices policy 269–70
cash 402, 421, 423
and the banking system 436–41
cash benefits 215–16
liquidity ratio, banks 459
money laundering 485
'cashless society' 83–84
'ceiling', exchange rate 581, 582, 583
central banks **442–43**
forward guidance 453–54
managing demand for credit 444–46
monetary policy 443–44
quantitative easing 449–53
transmission mechanism of interest rate policy 447
centrally planned economies, communist countries 9–10
Chang, Ha-Joon 4, 601–02
charitable giving, nudge theory 43
China
aid and infrastructure to Africa 603, 604
current account surplus 555
devaluation of currency 568
influence on UK inflation rate 410
purchase of US debt 568
'state capitalism' 10
'trade war' with USA 517, 530, 559
transfers of technology to 517
choice architecture **40**
default choice 40–41
choice(s) **14**
constraints restricting 29
framing 41
importance of the margin 29–30
mandated and restricted 42
need for 14
and opportunity cost 14–16
circular flow of income 330–31 *see also* investment; saving(s)
and balance of payments 554
equilibrium national income 331–32
government sector 332–33
national income 328–29
overseas sector 333
Phillips machine demonstrating 335
claimant count **302**, 303, 382
climate, link to economic performance 342
climatic cycles 378–79
closed economy **330**
Coase, Ronald
nature of the firm 82
private property rights 239
cognitive biases **37**
collective bargaining **182–83**
collective restrictive trading practices 269–70
collusive oligopoly 147–49
command (planned) economies **9**
resource allocation 9–10
commercial banks **432–33** *see also* retail banks
commercial bills **426**, 427, 441
commercialisation **498**, 506
commodity breakdown, UK trade 536–37
commodity money 419–20
commodity prices, effect on domestic inflation 408–09
common currency 586, 589
common currency areas 587–88, 589
common economic policies, EU 539
common fiscal policy 588, 589
comparative advantage **523–25**
assumptions underlying 526–27
and changing world trade patterns 533–34
David Ricardo's views 526
vs competitive advantage 527
vs strategic trade theory 531
competing supply 70, 71
Competition and Markets Authority (CMA) **266–67**
cartel agreements 269–70
mergers investigations 269
competition policy, UK **265**
mergers 269
monopoly policy 266–69
restrictive trading practices 269–70
competitive advantage
and import controls 528
and long-term direct capital flows 551
vs comparative advantage 527
competitive demand **70**
competitive devaluation 568
competitive markets **46**, 130
competitiveness
measure of, real exchange rate 572

supply-side policies 562–63
complementary goods **48**, 56, 70
complete market failure **228**
composite demand **70**
concentration ratio **145**
condition of supply **58–59**
conditions of demand **48–49**
and conditions of supply 58–59
externalities affecting 238, 247–48
technology reducing 114, 130
X-efficiency and X-inefficiency 138–39
confidence indicators 313
congestion charging 233–34, 249–50
constant returns to scale **89**, 526
consumer confidence 345–46
consumer expenditure 54
consumer goods **11**
consumer prices index (CPI) **304**
construction of 316–17
consumer surplus **159–60**
and price discrimination 161–62
and protectionist policies 531–32, 533
consumption
determinants 343–46
higher interest rates reducing 446
consumption externalities **237**, 238, 254–55, 257–58
consumption-led growth 563–64
contestable market theory **156–57**
and deregulation 276
sunk costs 158
contractionary fiscal policy **468**
contractionary monetary policy **445**
corporate bonds **427**, **429**, **430–31**
corporation tax 476, 478, 483
corruption **597**
cost-push inflation **397**, 399, 400, 401
and floating exchange rates 579
Phillips curve 413
and world commodity prices 408–09
council tax 478, 482, 485
coupon **428**
creative destruction **116**
credit **435**
availability of 346
creation of by banks 419, 435–38
demand for, central bank managing 444–46
credit crunch **346**, 348, 460–61, 551–52
cross elasticity of demand **56–57**
crowding out **469**, 492
financial 492, 495, 503
reversed by cuts in government spending 504
currency markets *see* foreign exchange markets
currency unions **586**
and economic integration 587
current account (of the balance of payments) **306**, **542**
applying *AD/AS* analysis to 554–55
four components of 543–45
imbalances, global implications 566–68
main sections 545–50
three factors influencing 557
UK's balance of payments 543–50
current account deficits **543**, 555–56
policies to cure or reduce 557–64
significance of 565–66
current account surpluses **543**, 556–57
policies to correct or reduce 564
significance of 565–66
customs union **538**
cyclical budget deficit **489**, 493
cyclical unemployment **387–88**, 389, 399
elimination of 468

D

de-globalisation 519–20
deadweight debt 494
deadweight loss **160**
and market failure 246, 248
debt **424** *see also* national debt
vs equity 424
decision making
biases in 37–38
by 'economic man' 35
importance of information for 31
prices coordinating 226
and rationality 24, 35–36
decrease in demand **48**
decrease in supply **59**
decreasing returns to scale **89**, 98, 526
default choice **40–41**
deficit financing **464–65**
Keynesian era 466, 487
deflation **304**, 395–96, 397
consequences of 407–08
expenditure-reducing policy 558–59
deindustrialisation **387**
demand **45**
demand for credit, Bank Rate 444–45
demand-deficient unemployment **387–88**
demand inflation **397–98**
demand for labour 164–71
demand-led spending 474
demand-pull inflation **397–99**
floating exchange rates contributing to 579–80
Phillips curve 413
demand side **367**
demand-side factors
affecting unemployment 389
and economic shocks 341
and long-run growth 369–71
short-run growth 367–8
demand-side fiscal policy **464**
using AD/AS diagrams to illustrate 467–68
demerit goods
and information failure 260–61
over-consumption of 257–58
tobacco and alcohol advertising 259–60
and value judgements 256–57
dependency theory of trade and development 513–14
deposit creation by banks 435–38
depreciation
of cars' value 73
of national capital stock 322
various meanings of 560
depreciation of currency *see* devaluation
deregulation 275–76, **498**
and contestable markets theory 276
of financial markets 507–08
and regulatory bodies 277
removing entry barriers 268, 506
derived demand **70**, **165**
devaluation 558
adjustable peg exchange rate system 583
competitive 568
effect on UK economic performance 584–86
exchange rate meaning of 559–60
expenditure-switching policy 559–60
and the J-curve effect 562
and price elasticity of demand 560
developing economies *see* less developed economies
digital downloads 71
diminishability (rivalry) **229**
direct controls 558, 559
direct investment, inward and outward 551
direct taxes **479**, 481–82 *see also* income tax
corporation tax 476, 478, 483
and household income 215–17
levied on capital and wealth 485
national insurance contributions (NICs) 219, 476, 478
'dirty floating' exchange rates 573, 583–84
disaster relief 602
discounting
books market 152
by supermarkets 146
discretionary fiscal policy **469**
discrimination *see also* price discrimination
in the labour market 191–95
diseconomies of scale **94–95**
external 98
internal 97–8
and long-run average cost curve 99–101
and returns to scale 98
super-tanker industry 101
disequilibrium **64**
disequilibrium trading 181

disinflation **304**, 397
disposable income 206, 217, 347, 475
disruptive innovation 114–15
distribution of income **197**
  benefits and costs of more equal 207–08
  consequences of government policies 222
  equality vs equity 204–05
  factors influencing 199–204
  'flat tax', case for and against 209
  income vs wealth 197
  measures of inequality 205–07
  and taxes 486
  World Bank classification of countries 198–99
distribution of wealth **197**
  factors influencing 203–04
division of labour **82–84**
divorce of ownership from control **125–27**
'double-dip' recession 368, 369
duopoly **116**

E

earnings/poverty trap **219–20**
economic activity 352
  aggregate demand and 353
  economic systems 8–10
  effects of government spending 471–75
  fluctuations in 375–77
  needs and wants 7–8
  and taxation 475–86
economic cycle **375–77**
  factors contributing to changes 377–79
  indicators for identifying phases of 379
  and output gaps 380
economic development
  vs economic growth 590–91
economic growth 299–300
  barriers to 597–600
  benefits and costs 372–73
  and the environment 374–75
  factors affecting 596–97
  policies to promote 600–01
  PPF diagram to show 18
  short-run and long-run 367–70
  theories of 371
  vs economic development 590–91
economic indices 304, 314–17, 319–20
economic integration 538–39, 587
economic liberalisation 10, 277 *see also* privatisation; deregulation
  marketisation 498, 506
economic methodology
  economics as a social science 2–3
  positive vs normative statements 5–7
  scientific methodology, use of in economics 3–4
  social vs natural science 4
economic models, building and using 334
economic and monetary union (EMU) 587
economic problem 14–15
economic resources 11–13
economic shocks **341**
  caused by the El Niño effect 342
  demand-side and supply-side 378
  and Keynesian fiscal policy 465
economic systems **8–9**
  price vs planning mechanism 9–10
  self-adjusting debate, Keynes 363
  UK as a mixed economy 10
economic unions 538 *see also* European Union (EU)
economic welfare **8**
  consumer and producer surplus 159–60
  deadweight loss 160
  effect of tariff/import duty on 533
  non-monetary benefits of working 172–73
  three components of 592–93
economies of scale **94–95**
  in car production 99
  external 98
  firm-level and plant-level 98
  internal 96–7
  internal vs external 95
  long-run average cost curve 99–101, 102
  and oil tankers 101
  and returns to scale 98
education and training **597**
  deprivation 213
  growth and development 596–97
  merit good 253, 254
  public spending on 472, 473
  state provision of 280, 285
effective demand **46**
efficiency 114
El Niño effect, economic shock caused by 342
elasticity **50–51**
elasticity of demand for labour **169–70**
elasticity of supply of labour **174–75**
employment *see also* unemployment
  and the economic cycle 379
  full employment 18–19, 301–02, 308, 389, 390, 391
  globalisation changing patterns of 517–18
  import controls protecting 529
  outsourcing of 201
  problem of under-employment 386
  short-term contracts 507
entrepreneurs
  factor of production 11–12, 77
  risk taking 110–11, 125–26, 508
entry barriers **121–22**
  and advertising 143–44
  deregulatory policies to remove 157, 268, 276
environmental issues
  as a scarce resource 12–13
  and economic growth 374–75
  Green Deal policy 242–43
  market failures 241
  and pollution 243–44, 282–83
equality **204**
equation of exchange 402–03
equilibrium **64**
equilibrium exchange rate 574, 575
equilibrium market prices, determination of 64–69
equilibrium national income **331**, 341
  and balance of payments 554–55
  demand-side fiscal policy 468
  effect of change in injections or withdrawals 331–32
equilibrium price *see* ruling market price
equilibrium unemployment **391**
equity (asset/wealth) **424**
  equity release 345
  negative equity 348
  positive and negative 424–25
  shares (equity market) 431, 442
  vs debt 424–25
equity (fairness) **204–05**
  horizontal and vertical 205
  and taxation 205, 481
  vs equality 204–05
European Union (EU) **515**, 586–89
  economic integration 539
  leave campaign, framing example 41
  and UK international trade 534–36
eurozone **299**, **539**, 586–87
  and immobility of labour 587–88
  lack of fiscal policy coordination 588
  volatile economic growth 299–300
excess demand **65**
excess supply **65**
exchange **83**
  equation of 402–03
exchange rate **571–72**
  effect of changes in on import price 561
  fixed 580–82
  freely floating 573–80
  and interest rate changes 455
  managed 573, 583–84
exchange rate equilibrium 574–75
Exchange Rate Mechanism (ERM) 584–85
exchange rate systems 570–86
excise duties 59, 476, 477, 478, 479, 483–84
excludable goods **229**
exit barriers **123**
expansionary fiscal contractionism (EFC) hypothesis 504

expansionary fiscal policy **468**
expansionary monetary policy **446**
expectations
  adaptive and rational 405, 416
  of future income and consumption 344
  of future inflation 346, 404–05
expenditure-reducing policy **558**
  deflation 558–59
expenditure-switching policies **558**
  devaluation 559–60
  direct controls 559
expenditure taxes *see* indirect taxes
expertise, as a form of aid 603
export-led growth **555**, 564
  and the UK economy 563
export subsidies **528**, 529
exports **306** *see also* trade
  effect of higher interest rates 446
external diseconomies of scale **95**
external economies of scale **95**
externalities **235–36**, 238
  causing market failure 246–50
  consumption 237
  demerit goods as negative 256, 257–60
  leading to wrong quantity of good/service 238
  merit goods as positive 253, 254
  negative, and allocative efficiency 250–51
  production 236–37

F

factor markets 164–65 *see also* labour market(s)
factors of production **11–12**, **77**
  and the costs of production 90
  distribution of national income between 199
  immobility of causing market failure 263–64
  and the law of diminishing returns 85, 86
  and returns to scale 89
fairness **39**
financial account, balance of payments **543**
  and capital flows between countries 550–52
Financial Conduct Authority (FCA) **458**
financial crisis (2007-08)
  bank failures 348, 440, 456, 458–59, 461
  causing temporary decline of financial services industry 548
  and collapse of tax revenues 485
  and the credit crunch 346, 348, 461
  negative equity created by 425
  recovery following 179
  and regulation of banks 508
  shifting the *LRAS* curve leftwards 360
  and systemic risk 434, 460–61
financial crowding out 492, 495, 503
financial markets **426–27**
  deregulation of 507–08
  supply-side policies 507–08
Financial Policy Committee (FPC) **457**
financial services industry, UK 548
  balance of trade in 547
  regulation 456–58
finite resources **12**
  depletion of 374, 375
firm-level economies of scale 98
firms **81**, 82
  government support to 350
  theory of the firm 77
firms' objectives
  divorce of ownership from control 125–26
  growth and survival 127
  profit maximisation 124–25
  sales revenue maximisation 127
  satisficing 128–29
fiscal drag 218
  and the poverty trap 219
fiscal policy **310**, **464** *see also* public expenditure; taxation
  and aggregate demand 464–68
  and aggregate supply 470
  and automatic stabilisers 490
  eurozone, importance of a common 588
  and the multiplier 469
fiscal watchdog 495–96
Fisher, Irving, equation of exchange 402–03
fishing industry
  mobile phone use, India 130
  and positive production externalities 237
  tragedy of the commons 240–41
fixed costs **90**
  vs sunk costs 158
fixed exchange rates **580–81**
  advantages and disadvantages 581
  government intervention 581–82
  rigidly fixed (managed) 583
flat taxes 209
flexible labour market, UK 80
floating exchange rates
  advantages of 576–77
  disadvantages of 578–80
'floor', exchange rate 581, 582, 583
fly-tipping 232, 287
'Footsie' 100 index 319–20
foreign direct investment (FDI) **551**
foreign exchange markets **426**, 431–32, 571
  central bank intervention in 582, 583
  and freely floating exchange rates 573–74, 575
  and speculative 'hot money' flows 578
forward guidance **453–54**
FPC (Financial Policy Committee) **457**
framing **41**
free-market economics **310**
  revival of, supply-side economics 497
  vertical *LRAS* curve 362
  view of budget deficits and surpluses 492–93
  vs interventionist supply-side policies 498–99
free-rider problem 230, 231, 232, 233, **235–36**
free trade
  benefits of 602
  fostering competitive advantage 527
  international agreements 540
  strategic trade theory 531, 602
  within the EU 539
free trade areas **538**
freely floating exchange rate **573–74**
  eliminating current account deficit 575
frictional unemployment **383**, 391
  reducing 393–94
  search theory 384
Friedman, Milton 34, 405, 415, 416
FSA (Financial Services Authority) 456
FTSE 100 index 319–20
full employment **19**, **301–02**
  Beveridge and 389
  macroeconomic policy conflicts and 308
  PPF diagram showing 18–19
  under-full-employment equilibrium, Keynes 362, 387, 403
full employment income **331**
fundamental economic problem **13–14**

G

GDP *see* gross domestic product
gender discrimination 192–95
General Agreement on Tariffs and Trade (GATT) 540
geographical immobility of labour **181–82**, 263–64, 383
  causing frictional unemployment 383
  and elasticity of labour supply 174
  ways of improving mobility 393–94
Gill, David, behavioural economics overview 35
gilts (gilt-edged securities) 425, 427–29
  and financial crowding out 492, 503
  retail banks investing in 442
Gini coefficient **205–06**
glass ceiling 194
globalisation **510–11**
  benefits and drawbacks of 514–15
  causes of 511
  consequences for LDEs 512–13
  consequences for MDEs 513
  de-globalisation 519–20
  and dependency theory 513–14
  main features of 511–12

migration of workers, effect on wages 201
and mobility of labour and capital 515–16
reducing power of national governments 516
role of MNCs in 516–19
in the service sector 515
glocalisation 512
GNI *see* gross national income
GNP *see* gross national product
good/benign deflation 407
Goodhart's Law 423–24
goods
balance of trade in 545–46
capital and consumer 11
complementary and substitute 48
government provision of public and merit 279–80
inferior and normal 49–50
merit and demerit 252–62
private, public and quasi-public 229–34
supply of 57–60
traded and non-traded 324–25, 326
'goods' and 'bads' 232
deflation 407
externalities 235
public 232
taxes 481
goods market 164–65
Google, abuse of market dominance 270
government bonds **427** *see also* gilts
Consols, redemption of 429
key characteristics of 428
and quantitative easing 449, 450–51
government failure **285**
examples of 286–7
reasons for 285–6
government intervention
exchange rate 581–82
and government failure 285–87
horizontal equity 205
interventionist supply-side policies 498–99
reasons for and against 278–79
to correct market failures 279–84
government spending *see* public expenditure
Great Depression (1930s) 300, 363, 466
Great Recession *see* financial crisis
Green Deal 242–43
gross (disposable) income 216–17
gross domestic product (GDP) **299** *see also* nominal GDP; real GDP
measured by output approach 329
per capita, as development indicator 592
vs gross national income (GNI) 322
gross investment 322, 349
gross national income (GNI)
and HDI ranking 595
as indicator of development 592
vs gross domestic product (GDP) 322
World Bank country classification 198–99
gross national product (GNP)
country comparison of living standards 324
inadequate indicator of economic development 569
indicator of living standards over time 323
and size of hidden economy 323

H

health
deprivation 213
healthcare, merit good 252, 253, 254, 256
insurance and moral hazard 261
public spending on 472, 473
hidden economy **323**
high street banks *see* retail banks
hit-and-run competition **158**
hoarding
of goods 406
of labour 80
of savings 332
horizontal equity 205
hot money **552**, 578, 582, 584–85
household income *see* income
household savings ratio 347–48
household spending 318–19
housing market 63, 181–82
house prices and wealth 345
human capital **321**, **599**
lack of preventing economic growth 599
new growth theory 371
Human Development Index (HDI) **593–95**
hypothesis of diminishing marginal utility **27**

I

ideas, new growth theory 371
immobility of labour **263**
contributing to market failure 263–64
and the eurozone 587–88
reason for wage differences 181–82
imperfect (asymmetric) information
effect on decision making 31
and frictional unemployment 383, 384
leading to market failure 262
imperfect competition 120, 121
Economics of Imperfect Competition (Robinson) 135, 178–79
of goods market, impact on MRP of labour 168
imperfectly competitive markets
monopsony labour market 178–82
and non-price competition 144
and wage differentials 181, 188
wage discrimination in 191–92
import controls
case for 527–31
direct controls 558, 559
GATT and WTO talks to reduce 540
quotas 528
removal of 564
import duties *see* tariffs
imports **306** *see also* trade
and protectionist policies 532, 533
trade patterns, UK 534–37
incentive function of prices **225**
incentives
prices creating 225
progressive taxation reduces 207–08
role of profit 110
supply-side fiscal policy 470
income
and consumption 344
earned vs unearned 200
effects of taxes and benefits on 215–17
inequality of distribution, measuring 205–06
World Bank classification 198–99
income distribution **346**
income elasticity of demand **55–56**
income tax 476, 477–78, 482–83
avoidance and evasion 482
Laffer curve 499–500
marginal and average rates 218, 480
supply-side views/policies 486, 505, 507
thresholds 218, 219
vs wealth tax 204
increase in demand **48**
increase in supply **59**
increasing returns to scale **89**, 98, 526, 528, 531
incumbent firms **122**
index numbers **314**
indexation **304**
indicators of development **592–93**
GNP figures providing inadequate 596
Human Development Index 593–95
indirect taxes **479**
and elasticity of demand 67
excise duties 476, 477, 478, 479, 483–84
VAT (value added tax) 478, 483–84, 485
vs subsidies 59–60
individual demand **47**
infant industries 528
inferior goods **49**, 55
inflation
causes of, changing theories over the years 401
expectations of future 346
target of 2% 303–04, 396, 397, 443, 444, 445
inflation psychology 404–05

informal (hidden) economy 221, 283, 286, **323**
information
  for decision making 31
  imperfect and asymmetric 31–32, 262
  imperfect market 181
  prices signalling 225
information failure **253**
  and demerit goods 255–56, 260
  merit goods 253, 260
infrastructure **598–99**
inheritance tax 478
inheritance and wealth 204
injections (into circular flow) **331**, 332–33
innocent barriers **122**
innovation **112**
  and creative destruction 116
  disruptive and sustaining 114–15
  profit as reward for 110–11
  vs invention 112–13
institutional factors **598**
interbank market 427, 441, 460–61
interdependence, oligopoly 147
interest rates **344**
  changes in, link to exchange rate 455
  and forward guidance 453–54
  inverse link to bond prices 428–29
  and monetary policy 443–46
  transmission mechanism 447
internal economies and diseconomies of scale **95**
international trade 520–21
  comparative advantage model 521–27
  economic integration 538–39
  and foreign exchange markets 431–32
  pattern of UK's 534–37
  protectionism 527–33
  uncertainty 578–79
  World Trade Organization 540
intervention of government in markets *see* government intervention
interventionist economists 279
interventionist policies **498**
invention **112**
  legal protection for (patents) 122
  vs innovation 112–13
inventories (stocks), effect of changes in 378
investment **331**
  in capital stock 321, 322
  determinants of 349–50
  higher interest rates reducing 446
  in human capital 599
  planned, circular flow model 332
  vs saving 349
investment banks **433–34**
  and the problems of systemic risk 434
investment-led growth 564
'invisible hand', of the market 129, 226
involuntary unemployment **382**

J

J-curve effect 562
Japanese manufacturing methods 81
Jevons, Stanley, sunspot theory of economic crises 378–79
John Lewis, employee-owned business 111
joint demand, complementary goods 56, 70
joint-profit maximisation, cartels 148–49
joint supply **69–70**

K

Kahneman, Daniel, slow and fast thinking 36
Keynes, John Maynard 311, 401
  belief in deficit financing 466
  and deficient aggregate demand 334, 382, 387
  General Theory 311
  inverted L-shaped *LRAS* curve 362
  on 'Poverty in Plenty' 363
  theory of consumption 344
Keynesian economists **309–10**
  cause of real-wage unemployment 391
  counter-cyclical demand management policy 490
  multiplier/accelerator interaction as cause of business cycles 378
  rejection of quantity theory of money 403
  theories of causes of inflation 401
Keynesian unemployment **387**
kinked demand curve model 149–50
Kodak 115, 116
Krugman, Paul 531, 587

L

L-shaped *LRAC* curve 102
Labour Force Survey (LFS) **302**, 303
labour immobility 181–82, 263–64
labour market(s) 164 *see also* unemployment
  demand for labour 164–71
  discrimination in 191–95
  imperfectly competitive 178–82
  national minimum wage 186–91
  perfectly competitive 175–78
  supply of labour 171–75
  supply-side measures 507
  trade union influence 182–86
  UK's 'flexible' 80
labour productivity **78**
  change in shifting demand curve for labour 169
  Japanese manufacturing methods 81
  productivity gap 78
  supply-side reforms increasing 370
  and technological change/progress 114, 371
  UK's productivity puzzle 80–81
  and wage/salary differentials 200
labour shortages, fruit-pickers **170–71**
Laffer curve 499–500
lag indicators 313
landfill tax 287
'law of contract' 598
law of diminishing marginal returns/marginal productivity **85–87**
law of diminishing marginal utility **27**
law of one price 130, 178
'law of unintended consequences' 286
lead indicators 312–13
leakages (withdrawals) from the circular flow 331–33
lean manufacturing 81
'lemons', market for 32
less developed economies (LDEs) **512**, 591–92
  barriers to growth and development 597–600
  dependency theory of trade 513–14
  globalisation consequences 512–13
liabilities **418–19** *see also* borrowing; debt
  in bank's balance sheet 436, 437, 439
life-cycle theory of consumption **344**
limit pricing 122, 123
Limits to Growth, The 374–75
liquidity **423**
  insufficient, and bank failures 458, 460–61
  retail banks 439–40
liquidity ratio **459**
liquidity trap 448
living standards
  changes in over time, NI data 323
  indicators of 329, 592–93
loans *see also* mortgages
  commercial banks 419
  hard and soft 602
  secured and unsecured 440
London interbank market 427
London Stock Exchange (LSE) 430–31
  FTSE 100 index 319–20
long run **53**
long-run aggregate supply (LRAS) curve **339–40**
  demand-pull inflation 397–98
  Keynesian (inverted L-shaped) 362
  position of 359–60
  supply-side fiscal policy 470
  and underlying economic growth 340
long-run average cost (*LRAC*) curves
  economies and diseconomies of scale 94–95, 99–101
  L-shaped, minimum efficient scale 102

long-run cost theory 77
long-run costs 90
long-run economic growth 18, **299**, 367
causes of 370, 371
trend growth rate 369–70
long-run equilibrium 133
long-run interest rates, bonds 428–29
long-run marginal cost (*LRMC*) 101
long-run Phillips curve 414–16, 501, 502
long-run production theory 77, 85
long-run profit maximisation
in monopolistic competition 141
in monopoly 136
in perfect competition 131–32
long-run supply 62
long-term (structural) unemployment **385**
Lorenz curve **205–06**

M

macroeconomic indicators **312–13**
man-made barriers to entry 122–23
managed exchange rates **573**, 583–84
managerial diseconomies of scale 96
managerial economies of scale 96
mandated choice **42**
margin 29–30
marginal cost of labour **176**, 177, 180, 185
marginal cost (MC) **91**
marginal physical product of labour **166**
marginal private benefit (*MPB*) 244–45
marginal private cost (*MPC*) 244–45
marginal propensity to consume **354**, 356
marginal propensity to save **354**, 356
marginal returns of labour **85**
law of diminishing returns 85–87
relationship with average returns 87, 88
marginal revenue (*MR*) **103**, 104–08
marginal revenue product of labour **167**, 177
marginal social benefit (*MSB*) 245
marginal social cost (*MSC*) 245
marginal tax rate **218**, 480
marginal utility **25**, 26–30
market **45**
market-based supply-side policies **498**
market conduct **147**
market cooperation vs collusion 147–48
market demand **47**
market demand curve 46–47, 48, 64, 66
for labour 168–69
market disequilibrium **64**
market economy **9**
role of profit 110–11
market equilibrium **64**
shift of demand disturbing 66
shift of supply disturbing 65–66
market failure **228–29**
correction of 279–84
regulation and 275, 280–81
market mechanism *see* price mechanism
market period supply 62
market power **136**
market structures **120**
competition dynamics 155–56
consumer and producer surplus 159–63
contestable markets 156–58
distinguishing between different 120–24
monopolistic competition 140–45
monopoly 135–40
non-contestable markets 156–58
objectives of firms 124–29
oligopoly 145–53
perfect competition 129–35
price discrimination 153–54
resource allocation 159
spectrum of 120
static and dynamic efficiency 159
market supply **57–58**
market supply of labour 171–72
factors shifting 173–74
monetary and non-monetary factors 172–73
marketisation 10, **498**, 506
Marshall–Lerner condition 561
Marxist explanation of economic cycles 378
maturity date, gilts **428**
maximum price laws (price ceilings) **283**, 286
merger policy 269
merit goods **252–53**
government provision of 279–80
information failure and 260
museums as 255
and positive consumption externalities 254–55
and uncertainty 261
under-consumption of 254, 275
and value judgements 256–57
minimum efficient scale (MES) 102
minimum legal prices (price floors) **283–84**
misallocation of resources *see* market failure
missing market **228**
mixed economy **9**
UK economy as 10
MNCs *see* multinational corporations
mobile phones *see also* Apple
Android operating system, Google 270
and market competitiveness 130
and opportunity cost 20
PPF diagrams 19
mobility of labour *see also* immobility of labour
improving 393–94
migration from poor to rich countries 515–16
within the eurozone 587–88
monetarism **401–02**
monetarists **401**, 402, 403
monetary policy **310**
Bank Rate 444–45
central bank implementing 443–44
contractionary 445
'conventional' and 'unconventional' 448
expansionary 446
and floating exchange rate 577
and forward guidance 453–54
interest rates 446, 447, 455
objectives and instruments 444
quantitative easing 449–53
and supply-side policies 502–03
Monetary Policy Committee (MPC) **444**
Bank Rate set by 445
monetary policy instruments **444**
money **419** *see also* money supply
as a medium of exchange 83, 419
as a store or value/wealth 419
characteristics of 419–25
defining 421
functions of 419, 421
vs barter 422
money GDP *see* nominal GDP
money illusion effect 415
money markets **426**
in the UK 427
vs capital markets 427
money supply **422**
Goodhart's Law 423–24
narrow and broad money 423
monopolistic competition 140–41
evaluation of 142
profit maximisation in 141, 143
monopoly **104**, **135**
average revenue and marginal revenue 105–07
competition policy 266–69
disadvantages of 138–39
dynamic efficiency 159
entry barriers protecting 121–22, 136, 138
and market failure 263, 275
natural 138, 139
possible advantages 137
privatisation increasing abuse 274
pure monopoly 121
short-run and long-run profit 136
X-inefficiency and 138–39
monopoly (abnormal) profit 136, 159
taxing 267
monopoly power **136**
monopsony **178–79**
effect of trade unions 184–85
wage rates and employment levels 180–81
monopsony power **179**

moral hazard **261**, **458**
more developed economies (MDEs) **513**
aid to LDEs 602–03
contract law 598
current account surpluses 556
strategic trade theory 528–29, 531, 601–02
terms of trade favouring 514
mortgages **197**
house price inflation 404
positive and negative equity 424–25
secured loans 440
sub-prime 460–61
motivational diseconomies of scale 97
multinational corporations (*MNCs*) **511**
and capital flows 550–51
impact on LDEs 512–13
and primary income 548–49
role in globalisation 516–19
multiplier **353**
and aggregate demand 353–54
automatic stabilisers reducing 490–91
dependent on nominal or real nation income 357
different types of 355
dynamic process 356–57
formula 356
interaction with accelerator 378
and Keynesian fiscal policy 466, 469
and marginal propensities to consume and save 356
size calculation 355
music industry
Apple's iPod 113
digital downloads 71

N

narrow money **423**
national capital stock **321**
depreciation of 322
national debt **487**
and the budget deficit 487–88
as % of GDP ('debt to GDP' ratio) 488, 494
interest payments on 475, 493
reproductive vs deadweight 494
rolling over 495
national defence, public good 231
national expenditure 329
vs aggregate demand 337
national income data, uses and limitations
to assess changes in living standards over time 323–24
to compare living standards between countries 324–25
national income multiplier *see* multiplier
national income (NI) **321–22**, 328–29
see also gross national income (GNI)
and the accelerator 350–51
and automatic stabilisers 491
equilibrium 331–32, 341, 554
and indicators of development 592–93
multiplier effects 353–54, 357
real vs nominal 329
shared between factors of production 199–200
national insurance contributions (NICs) 219, 476, 478
national living wage (NLW) 190, 219
national minimum wage (NMW) **186**
arguments for and against 187–89
effects on labour markets 186, 283
replaced by national living wage 190
national output **321**, 329
national product (NP) **321** *see also* national income (NI)
national wealth **321**
nationalised industries 271–72
nationalism, economic 518–19
natural barriers **122**
natural monopolies, utility industries 138, 139
natural rate of unemployment **391**
and Phillips curves 414–16
natural sciences vs social sciences 4
needs **7**
basic needs and poverty 210
vs wants 7–8
negative consumption externalities 237
negative economic growth *see* recession
negative equity 348, 424–25
negative externalities **235**
consumption 237
and demerit goods 256, 257–58
and living standards 324
policies to reduce 280–81
production 236
negative inflation (deflation) 304
negative output gap **380**, 381, 412
negative production externalities 236, 238
cause of market failure 246
neoclassical growth theory 371
net investment 322, 349, 351
new growth theory 371
nominal GDP **299**
and debt interest 475
national debt as a percentage of 488, 494
nominal vs real national income 329, 357
non-excludability (public goods) 230
and free-rider problem 235–36
non-interventionist supply-side policies **498**
non-monetised economy 323, 324
non-price competition 144, 155
non-renewable resources **12**
non-rivalry (public goods) 230–31
'normal capacity' level of output **359**, 361
and demand-pull inflation 397–98
and expansionary fiscal policy 468
and output gaps 380
and short-run growth 368
normal good **49**
and income elasticity of demand 55–56
normal profit **109**
in the long run 132
normative concepts
equity 204–05
fairness 39
normative statements **5**
nudges **38**, 42
automatic pension scheme enrolment 221
and choice architecture 40
vs shoves 42–43

O

occupational immobility of labour **181**
causing frictional unemployment 383
and elasticity of labour supply 174
occupational immobility of labour 263
ways of improving mobility 394
Office for Budget Responsibility (OBR) **472**, 495
as a fiscal watchdog 495–96
Office for National Statistics (ONS)
on construction of consumer price indices 316–17
'continuous revision' method of publishing economic data 368–69
income of poorest vs richest fifth 215–16
inequality in 1980s and 1990s 486
other economic indices constructed by 319
position on retail prices index (RPI) 305
under-employed workers 386
unemployment figures 303
wealth definitions 202
oligopoly 145–53
open economy **332**
opportunity cost **14–15**, 20
and rational behaviour/choices 15–16
optimal currency areas, UK and USA 587–88, 589
organ donation, and default choice 40–41
Ostrom, Elinor 240
output gaps **380**
and *AD/AS* gaps 380
and the economic cycle 380
link to inflation and unemployment 412–13
and UK economy 381
output per unit of capital employed 78, 93

output per worker *see* labour productivity
outsourcing of jobs 201

P

partial market failure **228**
  merit and demerit goods leading to 252
patents 122
  and drug companies 136
pensions
  private pension assets 203–04
  state pension and poverty 211–12, 214
  'triple lock' 472
  workplace schemes for low-paid workers 221
perfect competition 121, 129–30
  average revenue and marginal revenue 104–05
  and efficient allocation of resources 132–35
  profit and loss in 155
  profit maximisation in 131–32
perfectly competitive labour markets 175–78
personal savings ratio 347–48
personal tax allowances 218, 477–78
Phillips curves **413**, 414–16
Phillips machine, circular flow 335
physical capital 321, 599
physical vs financial assets 425
physical vs financial investment 349
planned economy **9**
plant **89**
plant-level economies of scale 98
policy conflicts **308–09**, 411–12
  and the Phillips curve 413–14
  reconciling 417
policy objectives **298–99**
  change in importance attached to 309–10
  conflicts and trade-offs between 308–09, 411–12
  economic growth 299–301
  and floating exchange rates 577
  full employment 301–03
  monetary policy 444
  price stability 303–06
  satisfactory balance of payments 306–07
political business cycle theory 377–78
political unions 538
pollution
  and economic growth 372, 373, 374
  negative externality 229, 235, 236, 241, 243–44, 246
  taxes 281
pollution permits 282–83
portfolio balance decisions 425
portfolio investment **551**
positive consumption externalities 237, 238
  and merit goods 254–55
positive economic growth 322, 376
positive economics 4
positive equity 424–25
positive externalities **235**, 238
  consumption externalities 237, 254
  merit goods 253
  production externalities 237, 247–48
positive output gap **380**, 381, 412–13
positive production externalities 237, 238
  causing market failure 247–48
positive statements **5**
poverty **210**
  absolute vs relative 210–11
  child poverty, attempt to redefine 211
  effects of, UK 213–14
  exploitation of low-paid workers increasing 188
  main causes of 211–12
  policies to alleviate 214–22
poverty/earnings trap 219–20
PPF (production possibility frontier) diagrams 17–22
predatory pricing 122, 123
price agreement **151**
price ceilings **283**, 284, 286
price competition 144, 155
price controls
  aviation industry 157
  ceilings and floors 283–84
  creating black markets 286
  to restrict monopoly abuse 267
price discrimination **153–54**
  conditions necessary for 154
  consumer surplus and 161–62
  limiting case 162
  and wage discrimination 192
price elasticity of demand **51**
  and devaluation 560
  effect of expenditure tax 67
  factors determining 52–53
  infinite and zero 51–52
  and monopolist's AR curve 107
  perfectly elastic demand 63
  and the slope of the demand curve 55
  total consumer expenditure 54
price elasticity of supply **60–61**
  factors determining 61–62
  housing market 63
  perfectly elastic supply 63
price floors **283–84**
price functions 225–26
  and market failure 228–29
price indices **304**
  consumer prices index (CPI) 316–17
  producer price indices (PPI) 319
price leadership **151**
price-maker **106**
price mechanism *see also* labour market(s)
  advantages and disadvantages 227
  allocating resources through 9
  extending, case for and against 227
  functions of prices 225–26
  and market failure 228–29
  and scarcity 13–14
price stability 303–05
  Bank of England's remit 442–43
price-taker **105**
price wars **151–52**
pricing strategies 122, 123, 144
primary income balance **543**
primary income deficit 545, 566
primary income flows 548–49, 550
primary (new issues) capital market 286, 426, 430, 431
principle of taxation **481**
private goods **229**
private pension assets 203–04
private property rights 235, 239, 599–600
privatisation **273**, **489**
  arguments for and against 273–74
  of monopolies 268
  regulation of privatised utility industries 277
  of UK water industry 139
pro-free-market economists 278–79, **310** *see also* free-market economics
  vs interventionist supply-side economists 498–99
producer surplus **160**
  and protectionist policies 532, 533
product differentiation **123–24**
  artificial barrier to entry 122
production **77**
  costs of 90–94
  methods of 113
production externalities **236–37**, 246–48
production possibility diagrams 17–22
production possibility frontier **17**
production theory 77
productive efficiency **21**, **114**
  and perfect competition 133
productivity **78**, 114 *see also* labour productivity
productivity gap **78**
profit **108–09**
  normal vs abnormal 109
  primary income 548–49
  and retail banks 439–42
  role of in a market economy 110–11
  taxation of 478
profit maximisation **108–09**
  in collusive oligopoly 148–49
  in monopolistic competition 141
  in monopoly 136
  objective of firms 124–25

in perfect competition 131–32, 155
vs revenue maximisation 127–28
vs satisficing 128–29, 138
profit-maximising rule (*MR* = *MC*) 125
progressive taxation **207**, 479, 480
property rights
lack of in less developed countries 599–600
patents providing IP rights 136
trading of private, Coase theorem 239
proportional taxation **480**
'flat tax' 209
protectionism *see also* import controls
case for 527–30
causes and consequences of adopting 531–32
strategic trade theory, Krugman 531
of the USA 556, 567
Prudential Regulation Authority (PRA) **458**
public expenditure 471
and the circular flow model 332–33
demand-led 474
national debt interest payments 475
social protection, health and education 472–73
total managed 472
transfer payments 474–75
public goods **230–31**
government provision of 279
public ownership **271–72**
purchasing power parity (PPP) exchange rates **325**
vs market exchange rate 325–26

Q

quantitative easing (QE) **449–50**
and the Brexit vote 452
(in)effectiveness of 451–52
five phases of (2009 to 2020) 450–51
inquiry into (2021) 453
government spending percentages 474
interpreting data presented in index number form 79
percentage calculations 22
plotting data on a bar chart 217
statistical data 22
quantity-setter **106**
quantity theory of money **402**
Keynesian economists' rejection of 403
and monetarism 401–03
quasi-public goods **233**
quintiles and quantiles 198
quotas **528**

R

rail industry
allocative efficiency and rail fares 251
renationalisation of East Coast rail 272
trade union action 185–86
trading private property rights, Coase theorem 239
rate of interest *see* interest rates
rational behaviour **24**
bounded rationality and self-control 35–36
and opportunity cost 15–16
rational expectations theory 405, 416
rationing function of prices **225**
real exchange rate 572
real GDP **299**, 314
changes in, economic cycles 379
macroeconomic indicator 312
vs nominal GDP 299, 300
real national income
effect of increase in aggregate demand 358
indicator of economic performance 329
vs nominal 329, 357
real rate of interest, calculating 344, 455
real-wage unemployment **390–91**
real wages **390**
and inflation 406
and natural rate of unemployment 391, 501–02
recession **299**, 376–77
and deficient aggregate demand, Keynes 334
definition by ONS 368–69
Great Depression (1930s), US 300
Great Recession (2008-2009) 585
and inventory changes 378
and monetary policy 445, 448
and UK jobs market 179
and UK labour productivity 80
recycling 12
and pricing of rubbish collection 232
redistributive policies 222
reflationary policies **336**, 397, 468, 564
reflective thinking 36–37
regressive taxation **479**, 480
excise duties 478
regulation **274–75**
advantages of 275
disadvantages of 276
of the financial system 456–58
and market failure 275
of privatised utilities 277
rate of return 268
to correct market failure 280–81
regulatory capture **276–77**
regulatory (watchdog) agencies 277
Competition and Markets Authority (CMA) 266–67
Office for Budget Responsibility (OBR) 495–96
relative poverty **211**
causes of in the UK 211–12
effects of in the UK 213–14
renewable resources **12**
representative money 420
researchers' ideas, new growth theory 371
resource allocation **19**
and dynamic efficiency 159
improved by floating exchange rate 576
and perfect competition 133
price vs planning mechanism 9–10
and profits 110
resources 12–13
depletion 375
scarcity of 13–14
restricted choice **42**
restrictive trading practices policy 269–70
retail banks **432–33**
conflicts between liquidity, profitability and security 440–42
deposit and credit creation 435–38
liquidity–profitability trade-off 439–40
structure of balance sheet 438–39
retail prices index (RPI) **304**
retained profits 111
returns to scale **89**
and comparative advantage 526
constant, decreasing and increasing 89
and economies/diseconomies of scale 98
revaluation 564
revenue 103–08
reverse causation 403
reverse J-curve effect 564
Ricardo, David 526
risk
entrepreneurs taking 12, 110–11, 125–26, 508
and internal economy of scale 96–97
systemic 434, 457, 460–61
rival goods **229**
roads/road use
charging for 249–50
congestion charge 233–34
quasi-public good 233, 249
vs rail 251
rubbish (garbage)
disposal, charging for 232
unlawful dumping of 287
rules of thumb **37**
ruling market price **46**
determination of 64–68
and perfect competition 104–05, 131, 132
ruling market wage 176–77, 178

S

salary differentials 175, 200–01
sales revenue maximisation 127
satisficing 36, **128–29**
saturation advertising 144

saving(s)
determinants of 347
higher interest rates encouraging 446
household savings 347–48
and income level 344
personal savings ratio 347
planned saving and equilibrium national income 331–33
policies to encourage 508
vs investment 349
Say's Law 388
scarce resource allocation *see also* resource allocation
allocative function of prices 226
fundamental economic problem 13–14
price vs planning mechanisms 9–10
scarcity **13–14**
diamonds and water paradox 28
of environmental resources 12–13
and need for choice 14, 15, 19
of positional goods 50
Schumpeter, Joseph 116
scientific methodology 3–4
search theory of unemployment 384
seasonal unemployment **389**
second-best theory 529
second-hand cars 31–32, 73
'second-hand' market, bonds and shares 427, 428, 430, 431
secondary income balance **543–44**
secondary income flows 549
secondary markets
black markets 283, 286
'second-hand' capital markets 430, 431
security (of loans) **440**
self-interest 130
services/service sector
balance of trade in 547
globalisation 515
shadow (informal) market 283, 286
shares **424**, 425, 431, 442, 551
capital gains and losses 429
shocks, economic *see* economic shocks
shopping basket, CPI 316, 317
short run **53**
short-run average total cost (*SRATC*) curves 95, 99–100, 102
short-run costs 90–93
short-run economic growth 18
demand side and supply-side determinants 367–68
short-run Phillips curve 413
and economic policy 415–16
link with long-run Phillips curve 414–15
short-run supply 62
short-term speculative 'hot money' capital **552**
floating exchange rates vulnerable to 578
signalling function of prices **225**
single market 589
smartphone technology
Apple and its competitors 117, 123–24, 156
impact on traditional camera industry 115
reducing barriers to labour market entry, Uber 264
Smith, Adam
absolute advantage 526
diamonds and water paradox 28
division of labour 84
four principles of taxation 481
'invisible hand' of the market 129, 226
Smith, David, Economic Outlook column *Sunday Times* 313
smoking *see* tobacco
social benefit 245, **253**
social cost **255**
social norms **38**
social protection/social security, public spending on 472, 473, 475
social science, economics as 2–3
specialisation **82–83**
and absolute advantage 522–23
and comparative advantage 524–27
dangers of over-specialisation 529
demotivating workers 97
and managerial economies of scale 96
speculative bubbles 377, 564
speculative buying and selling, bonds 429
speculative demand 49
speculative 'hot money' capital flows 552, 578
stamp duty 478
standard of deferred payment, money as 421
state intervention *see* government intervention
state pension 211–12, 216, 305, 472
static efficiency **159**
stock exchange 430, 431
FTSE 100 index 319–20
strategic (artificial) entry barriers **122**
strategic trade theory 528–29, 531, 601–02
counter-argument 602
structural budget deficit **489**
benefits and costs of 493
structural unemployment **385**
equilibrium unemployment 391
international competition as cause of 387
policies to reduce 393–94
sub-prime mortgages 346, 348, 460–61
subsidies **254**
on exports 528, 529
R&D spending 506
to correct market failures 280
to museums 255
vs expenditure taxes 59–60
substitute goods **48**
competitive demand 70
and cross elasticity of demand 56
perfect substitutes 63, 105, 107, 140–41
and price elasticity of demand 52, 53
sunk costs **122**
sunset industries 528
sunspot theory of economic crises 378–79
supermarket industry
competition in 146
exchange rate changes and import prices 561
supernormal profit **109**
supply **45**
supply of labour 171–75
supply side **368**
supply-side economics **496–98**
influence on UK politicians 504–05
supply-side factors
affecting unemployment 389
causing economic shocks 341
leading to short-run economic growth 368
triggering economic cycles 377
supply-side fiscal policy **470**
and national debt reduction 487
widening inequality 486
supply-side improvements **497–98**
supply-side policies **470**, **496**
and the current account 562–63
free-market vs interventionist 498–99, 508
inflation and balance of payments 503
Laffer curve 499–500
microeconomic 505–08
and monetary policy 502–03
role in reducing natural rate of unemployment 501–02
to reduce unemployment 394, 414, 416
using *AD/AS* diagram to show 470
survival, business objective 127
systemic risk **434**, 457, 460, 461

T

tariffs
case for 527–30
and economic welfare 533
free trade area vs customs union 538
GATT and WTO agreements 540
tax credits 188–89
tax revenue 475–76, 485
from excise duties 477
Laffer curve 499–500
tax thresholds 218, 219
taxation **281**
average tax rate 480, 499–500
on capital and wealth 204, 476, 485
direct 477–78, 479, 482–83
and distribution of income and wealth 486

indirect 476, 477, 478, 479, 483–84
marginal tax rate 218, 480–81
of monopoly profit 267
principles of 481
progressive 479, 480
proportional (flat) 209, 480
regressive 479–80
role and merits of different 481–82
taxi market, London 234, 264
technical economies of scale 96
indivisibilities 122
technical progress **17**, 112
causing demand curve for labour to shift 169
and economic growth 340, 370
influencing investment decisions 350
sources of, new growth theory 371
technological change **112**
allowing electronic road pricing 233
and creative destruction 116–17
digital downloads vs CDs and DVDs 71
disruptive innovation 114
effect on economic performance 113–14
invention vs innovation 112–13
and the structure of markets 116
sustaining innovation 114
technological unemployment 385
technology transfers 517
terms of trade 514
'three Rs' of reflation, removal of import controls and revaluation 564
'tied' aid 603
time period
and price elasticity of demand 53
and price elasticity of supply 62
tobacco
advertising 259–60
demerit good 255, 256
negative externalities 257–58
price elasticity of demand 53
taxation of 53, 282, 479
token money 421
total costs **90–91**
and profit 108–09
total managed expenditure (TME) **472**
total net wealth 202
total returns **86–87**
total revenue **58**, **103**
relationship to marginal revenue 108
'toxic assets/debts' 461, 551–52
trade **83**, 520–21
comparative advantage model 521–27
effect of tariffs on economic welfare 533
import controls and protectionism 527–31
pattern of UK's trade 534–37
pattern of world trade, change in 533–34
promoting growth and development 601–02
protectionism, causes and consequences 531–32
trade cycle *see* economic cycle
trade liberalisation *see* free trade
trade-off between policy objectives **308**, 411
trade unions **182–83**
trade war, China-USA 517, 530
traditional vs behavioural economics 34, 36, 40
'tragedy of the commons' 239–40
and the fishing industry 240–41
training *see* education and training
transfer payments 474–75
vertical equity 205
transitional unemployment **383**
transmission mechanism of interest rate policy 447
Treasury bills **426**, 427, 441, 502–03
trend growth rate **369–70**
'trickle-down' effect 504
'triple lock', state pension 212, 472

U

Uber taxis 264
Ukraine war, supply shock 378
uncertainty
Covid-19 pandemic and economic growth 299–300
effect on household savings 347–48
international trading 578–79
in oligopoly 147
and under-consumption of merit goods 261
under-employment 386
under-full-employment equilibrium 403
underground (hidden) economy 221, 286, **323**
unemployment **19**
consequences of 392–93
cyclical 387–88
demand-side and supply-side factors 389
equilibrium 391
frictional 383–84
government policies to reduce 393–94
and inflation, Phillips curves 413–16
measures of 302–03, 382
natural rate of 391
and poverty 212
PPF diagrams illustrating 18–19
real-wage theory of 390–91
reducing natural rate of 416, 501–02
seasonal 389
structural 385–87
and trade unions 183
voluntary and involuntary 382
unemployment trap 220–21
United Nations Human Development Index (HDI) **593–95**
United States (USA)
capital flows from China 568
current account deficit 567–68
de-globalisation 519–20
Great Depression (1930s) 300
influence on UK inflation rate 410
'McDonaldisation' effect 512
nationalism and protectionism 518–19, 529–30, 567
sub-prime mortgages 460–61
technology transfers 517
Universal Credit 220, 392
vs means-tested benefits 215
utility **25**
law of diminishing marginal 27
maximisation 28–29
revealed preference theory measuring 30
total and marginal 26–27
utility industries
natural monopolies 139
rate of return regulation 268
regulation of privatised 277

V

value judgements 5
fairness and equity 39, 204
influencing decision making and policy 5–6
merit and demerit goods 256–57
variable costs **90**
VAT (value added tax) 59, 478, 483–84, 485
Veblen goods 50
velocity of circulation of money 402, 403
vertical equity 205
volume economies of scale 96, 101
voluntary unemployment **382**
vs non-rivalry, public goods 230–31

W

wage-cost inflation 408
wage differences 175
reasons for 181–82
and supply and demand theory 200–01
wage discrimination **191**
condition necessary for 192
gender pay gap 192–95
and price discrimination 192
'wage floor' *see* national minimum wage (NMW)
wage stickiness 390–91
wages
poverty and 212
real-wage unemployment 390–91
vs non-monetary considerations 172–73
wants **7–8**
wealth **197**
distribution of, factors influencing 203–04

influence on consumption 345
money as a store of 419
national 321
'old' vs 'new' 204
portfolio 425
taxation 204
welfare benefits
benefit cap 392
choosing in preference to low pay 486
indexation of 304
and long job search periods 384
reducing to incentivise working 498, 507
state pension 211–12, 216, 305, 472
transfer payments 474–75
unemployment 212, 384, 474
Universal Credit 220
universal vs means-tested 215, 219
withdrawals (from circular flow) **331**, 332–33
workplace pension schemes 221
world trade, changing patterns of 533
World Trade Organization (WTO) **511**, 540

X

X-efficiency/inefficiency 138–39
Xerox 115

Y

yield (of a bond) **428–29**

Z

zero-hours employment contracts 80, 386, 507
zero inflation 303, 308, 405, 406, 414
zero lower bound, Bank Rate floor 448
zero price elasticity of demand 51–52
'zombie' firms 80

AQA A-level Economics: Boost

Created with teachers and students in schools across the globe, Boost is the next generation in digital learning for schools, bringing quality content and new technology together in one interactive website.

AQA A-level Economics Fifth Edition: Boost Core includes a subscription to Boost where you will find a range of online resources to support your teaching.

- **Engage.** Lessons and homework brought to life with inventive presentations, worksheets, case studies and 'stretch and challenge' material to incorporate into your own trusted resources.
- **Plan.** Saves time planning and ensures course coverage with a course planner accompanied by expert author guidance.
- **Support.** Assessment for learning with comprehensive knowledge tests and new examples of exam-style questions and answers.

AQA A-level Economics Fifth Edition: Boost eBook

Boost eBooks are interactive, accessible and flexible. They use the latest research and technology to provide the very best experience for students and teachers.

- **Personalise**. Easily navigate the eBook with search, zoom and an image gallery. Make it your own with notes, bookmarks and highlights.
- **Revise**. Select key facts and definitions in the text and save them as flash cards for revision.
- **Switch**. Seamlessly move between the printed view for front-of-class teaching and the interactive view for independent study.

To subscribe or register for a free trial, visit
https://www.hoddereducation.co.uk/aqa-economics-alevel-boost

The Boost course and Boost eBooks have not been entered into the AQA approval process.

Boost

AQA
A-level

Economics

Fifth Edition

This title has been approved by AQA

Key topics for AQA A-level Economics, revised with new material including refreshed case studies, exam-style questions and study tips.

- Provides strong support in preparing for the topics on the AQA A-level Economics specification
- Features up-to-date case studies that engage with the latest economic developments, including the effects of Brexit, the pandemic and more
- Includes features to help with quantitative skills support and developing chains of reasoning, and making links between topics, with updated exam-style questions
- Gives relevant background and real-world examples to help understanding of key economic issues, with clear and simple explanations of core concepts
- A new edition of a trusted resource, which maintains and develops its accessibility to make economics compelling for a wider audience

Ray Powell is an experienced author and teacher with examining experience. He has written numerous best-selling textbooks on AQA Economics.

James Powell is an experienced Head of Economics and author.

David Horner is an experienced teacher of Economics, author and senior examiner.

Steve Stoddard is an experienced Head of Economics, author and senior examiner.



Boost

This title includes an eBook and digital teaching and learning support.

Visit hoddereducation.co.uk/boost to find out more.

Answers are available at:
www.hoddereducation.co.uk/aqa-economics-alevel

HODDER EDUCATION
t: 01235 827827
e: education@hachette.co.uk
w: hoddereducation.co.uk

ISBN 978-1-3983-7519-2

FSC