CAMBRIDGE
UNIVERSITY PRESS

Economics
for the IB Diploma

COURSEBOOK

Ellie Tragakes

Third edition

Digital Access

Cambridge Panel
Together with IB teachers

CAMBRIDGE
UNIVERSITY PRESS

Economics

for the IB Diploma

COURSEBOOK

Ellie Tragakes

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit
of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108847063

First published 2009
Second edition 2012
Third edition 2020

20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2

Printed in Malaysia by Vivar Printing

A catalogue record for this publication is available from the British Library

ISBN 978-1-108-84706-3 Coursebook with Digital Access
ISBN 978-1-108-81065-4 Digital Coursebook
ISBN 978-1-108-81066-1 Coursebook - eBook

Additional resources for this publication at www.cambridge.org/9781108847063

Contents

For

Alexios, Alkeos, Kyriakos

About the author

Ellie Tragakes has a BA from Columbia University, MSocSc from the University of Birmingham and PhD from the University of Maryland. She has worked in the areas of economic development of agriculture in economically less developed countries, financial services and health systems, in several national and international organizations, including the World Bank and World Health Organization. She is a highly experienced author, with numerous professional publications. Many of these are in the area of health systems financing and reforms in transition economies, and have been translated into several languages including Russian and Chinese. She is also a highly experienced teacher and examiner, having taught for many years in the Economics Department at the American College of Greece and also having served as IB Economics Chief Examiner. She currently serves as IB Senior Examiner and Managing Director of Hellenic Agricultural Enterprises.

Acknowledgements

Author

I would like to express my deepest gratitude to Peter Rock-Lacroix for his detailed, thorough and exhaustive review of the entire book, including the materials accompanying the digital version. In addition to offering most insightful and creative suggestions for improvements and catching errors, Peter was a continuous encouraging and supportive presence throughout the entire writing process.

I am also grateful to Dimitris Doulos, Roma Kaur and Charles Wu for their valuable comments on the first three chapters of the book that helped set guidelines for the remainder of the writing.

In addition, I would like to extend my sincere thanks to many friends and colleagues around the world who contributed to the previous two editions of the book. They include Henry Tiller, former IB Economics Chief Examiner, and Emilia Drogaris, both of whom commented extensively on the second edition, and Julia Tokatlidou for her extensive review of the first edition. I would also like to express my gratitude to Tibor Cernak, Simon Foley, Hana Abu Hijleh, Kiran Asad Javed, Jane Kerr, Pat Lasonde, James Martin, Peter Rock-Lacroix, Sachin Sachdeva, Vijay Peter D'Souza, Charles Wu, Lar Lun, Constantine Ziogas for their comments and suggestions for improvements.

My warmest thanks also go to K.A. Tsokos for the guidance and inspiration that his book, *Physics for the IB Diploma*, provided for me.

Author and publisher

The author and publishers acknowledge the following sources of copyright material and are grateful for the permissions granted. While every effort has been made, it has not always been possible to identify the sources of all the material used, or to trace all copyright holders. If any omissions are brought to our notice, we will be happy to include the appropriate acknowledgements on reprinting.

UN Sustainable Development Goals from https://www.un.org/sustainabledevelopment/sustainable-development-goals/ © 2019 United Nations, reprinted with permission of the United Nations

Thanks to the following for permission to reproduce images:

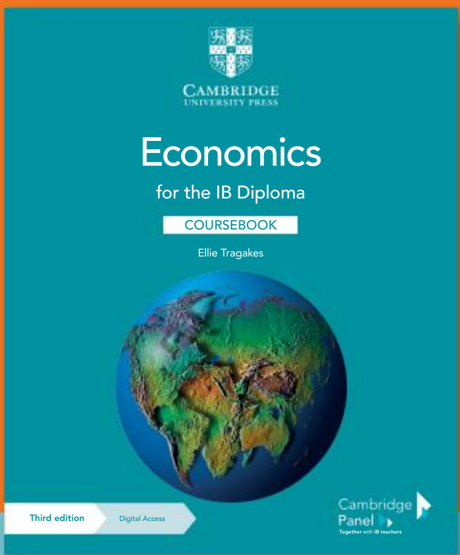
Cover Russ Widstrand/Getty Images

Inside Tony Wu Photography/Shutterstock; Hulton Archive/GI; Culture Club/GI; Bettmann/GI; BSIP/GI; Florian Gaertner/GI; Miroslav_1/GI; Jaboo2foto/GI; Ullstein Bild/GI; Rawpixel.Com/Shutterstock; Artmarie/GI; Craig Ferguson/Lightrocket/GI; Luoman/GI; Carsten Koall/GI; Quangpraha/GI; Douglas Graham/GI; Lillitve/GI; Florian Gaertner/GI; Maja Hitij/GI; Education Images/GI; EtiAmmos/GI; Per-Anders Pettersson/GI; Ofirperetz/GI; Stockshoppe/Shutterstock; Tommasourbinati/GI; John Thys/GI; John Longley/GI; Jasondoiy/GI; Jasper Juinen/GI; Pavelvinnik/GI; Str/GI; Tony C French/GI; Pongkiat Rungrojkarnka/GI; Yuliya Derbisheva/GI; Aliraza Khatri's Photography/GI; Petmal/GI; Senhan Bolelli/Anadolu Agency/GI; Flyingrussian/GI; Dszc/GI; Hadynyah/GI; Oli Scarff/GI; Drante/GI; Kheng Guan Toh/Shutterstock; Smith Collection/Gado/GI; John Moore/GI; Blocberry/GI; Brozozowska/GI; Dea/G. Dagli Orti/GI; D3sign/GI; Tim Graham/GI; Robert Hradil/GI; Rebeca Mello/GI; Australian Scenics/GI; Benoitb/GI; Str/GI; Barcroft Media/GI; Ted Aljibe/GI; The Washington Post/GI; Frankvandenbergh/GI; Tkkurikawa/GI; Jxfzsy/GI; Maria Toutoudaki/GI; Pacific Press/GI; A_Noina/Shutterstock; Farm Images/GI; Ute Grabowsky/GI; Himarkley/GI; Jan Sochor/GI; Moment/GI; Brazil Photos/GI; Guido Dingemans, De Eindredactie/GI; Sopa Images/GI; Karelnoppe/GI; Nurphoto/GI; Jasmin Merdan/GI; John Borthwick/GI; Ashraf Shazly/GI; Bennett Raglin/GI.

Key: GI= Getty Images.

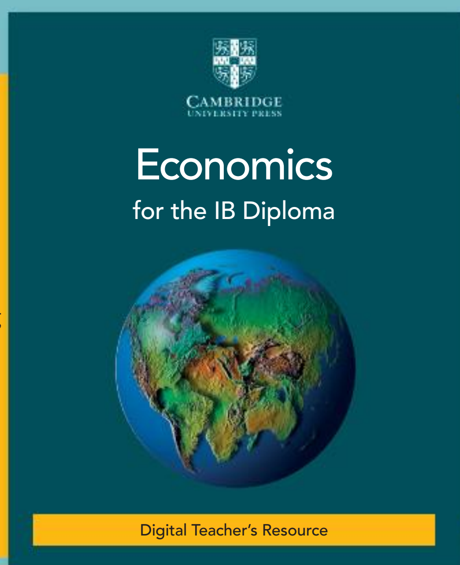
How to use this series

This suite of resources supports students and teachers of the Economics course for the IB Diploma programme. The coursebook and the teacher's resource work together to support teachers and students on their learning journey, providing the necessary knowledge and skills required to succeed in the course.



The coursebook with digital edition provides full coverage of the latest IB Economics guide. It includes activities and exam-style questions that test students understanding and develop problem-solving skills, links to TOK and real-world examples of economic principles. With clear language and style, the coursebook is designed for international learners.

Developed in collaboration with IB Economics teachers from the Cambridge Panel, this teacher's resource provides you with answers to the activities in the coursebook and to the exam-style questions/papers. It also includes advice on sourcing and using case studies, EAL and subject-specific vocabulary support, evaluative essay writing, exam practice, presentations, extra activities, and downloadable worksheets.



How to use this book

Throughout this book you will find lots of different features to help your learning.

CONCEPTUAL UNDERSTANDINGS

These statements introduce the key concepts of the unit.

LEARNING OBJECTIVES

A bulleted list at the beginning of each section clearly shows the learning objectives of the section. These objectives link to the assessment objectives and outcomes in the IB economics guide.

TEST YOUR UNDERSTANDING

Test your understanding questions appear at the end of every topic, and help you review the learning objectives of the section. They can be used as the basis for class discussions or homework assignments. If you can answer these questions, it means you have understood the important points of a section.

REAL WORLD FOCUS

Real word focus features help you relate the theory you are learning to practices in real life. These are followed by questions intended to focus your attention on important theoretical ideas and their relevance to real world situations.

Key points, such as important laws, concepts, definitions and conclusions, are highlighted in an orange box. This helps you focus on the important material in a chapter and can facilitate reviewing.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

BEFORE YOU START

This short section of statements and questions will help you to reflect on prior learning, check what knowledge you will need for the chapter and provoke your own thoughts about the topic.

Division between core and higher level material

The book is divided into core and higher level material. A vertical line runs down the margin of all higher level material, allowing you to easily distinguish higher level from core material.

THEORY OF KNOWLEDGE

Theory of knowledge features encourage you to think critically about economics as a social science, the nature of economic knowledge, difficulties involved in acquiring economic knowledge, why economists disagree, the role of values, language, ethics, beliefs and ideology in the development of economic knowledge. Each feature ends with questions intended to stimulate further thinking and discussions on these important theory of knowledge issues.

INQUIRY AND REFLECTION

These questions appear at the end of each chapter and encourage you to reflect on the development of your skills proficiency and your progress against the objectives. They are intended to encourage your critical thinking and inquiry-based learning.

Key terms are highlighted in **orange bold** font at their first appearance in the book so you can immediately recognise them. At the end of the book, there is a glossary that defines all the key terms.

Note to the reader about the digital coursebook

The digital coursebook contains a number of files that are intended to assist and guide you as you use this coursebook – all editions of the book come with access to these documents. For those using the print version, instructions on how to access the digital coursebook can be found on the inside cover. The digital coursebook files include the following:

Introduction to IB Economics assessment: Exam papers and internal assessment

Here you will find an explanation of all the IB exam papers you will need to prepare for, including information on learning objectives with corresponding assessment objectives that appear at the beginning of each section and subsection of the book, command terms that appear at the beginning of exam questions, the duration of each exam paper and the percentage of each in the total IB score, useful suggestions to help you with exams, and an explanation of internal assessment.

Checklists of important topics to help you organize the coursebook materials

Checklist: Real world issues and key concepts

Each of the three main units of the book (Microeconomics, Macroeconomics and The Global Economy) is divided into two sections headed by a *real-world issue*. There are six real-world issues, each focusing on a broad, general question introducing the main topics of the subsequent chapters. In addition there are nine *key concepts* running through your economics course. In this checklist you will find suggestions on how to link the key concepts to the real-world issues.

Checklist: Measures of economic activity, well-being, economic inequality and poverty

There are several partly overlapping measures of these important economic variables. They are all listed here with a summary definition of each, to help you distinguish between them and avoid confusion.

Checklist: Calculations at SL and HL

The IB Economics syllabus provides detailed information on all calculations that you should be able to perform, at both SL and HL. This checklist lists all these calculations so that you can easily go through them to make sure that you understand what they are and check them off as you go through practice exercises. You will find numerous practice exercizes in the Test Your Understanding questions in the book.

Checklist: Policies for Paper 3 (HL only)

One of the questions in Paper 3 asks students to recommend a policy in order to address a particular issue. This checklist includes all policies that are included in the IB Economics syllabus, along with the chapter where the policy is discussed. By going through this list you can easily check off that you are familiar with all the policies that you may be required to recommend in Paper 3. You will find numerous questions relating to policies in the Test Your Understanding questions as well as in Inquiry and Reflection activities. In addition your teacher may give you policy questions from the teacher's resource.

Important diagrams with tips on how to use them

The IB Economics syllabus provides a detailed list of all diagrams that you need to understand and draw. Here you will find all these diagrams reproduced and organized by chapter and topic within each chapter, with a note if the diagram is HL only. In addition for each diagram there is a tip on what you should be able to illustrate by use of the diagram in question. Your teacher can give you practice questions with diagrams from the teacher's resource.

Quantitative techniques

This section contains all the quantitative techniques you need to understand in order to excel in your IB Economics course. It enables you to review everything from percentages and percentage changes to understanding the essentials of relationships between variables, and interpreting and constructing diagrams and graphs.

Paper 1 (HL and SL)

Here you will find sample questions for this paper at SL and HL, corresponding to each chapter in the book, covering all the chapters. Your teacher can provide you with markschemes that are included in the teacher's resource.

Paper 2 (HL and SL)

Here you will find two complete Paper 2 questions. Going through these carefully will provide you with a good idea of the structure of this paper. Your teacher can provide you with markschemes that are included in the teacher's resource.

Paper 3 (HL only)

In this part you will find two complete Paper 3 questions. Going through these carefully will allow you to understand how this paper is structured. Your teacher can provide you with markschemes that are included in the teacher's resource.

Supplementary material

This is an extension of the material in the book. It includes a number of topics that are not required material, but that may be of interest to students who would like to gain a deeper understanding of some issues in economics.

Theory of Knowledge features

Real World Focus features

Unit 1
Introduction to Economics

CONCEPTUAL UNDERSTANDINGS

1 Economics is a social science in which there is ***interdependence*** between people who interact with each other to improve their ***economic well-being***, and who are influenced and empowered by their values and natural surroundings.

2 The economic world is dynamic and subject to continuous ***change***.

3 Economic theories are founded on the principles of logic complemented by empirical data, used to build models that try to explain our complex economic world. Individuals and groups of individuals display motivations and behaviours that are complex and varied. Their understanding is facilitated by the contribution of several disciplines that interact with economics, such as psychology, philosophy, politics and history.

4 Economic decision-making is central to determining the relative ***economic well-being*** of individuals and societies.

5 The key problem of economics derives from ***scarcity,*** which necessitates ***choice***. This in turn leads all economies to confront trade-offs, opportunity costs and the challenge of ***sustainability***.

6 Economic thinking faces a number of debates. Among the most important of these is the debate between economic growth and ***equity***, and the debate between free markets (and their ***efficiency***) and government ***intervention***.

7 Unlimited economic growth based on the use of finite resources cannot continue indefinitely. Established theories and approaches to the issue of economic growth are being challenged by new models and social movements with a view to redesigning the economy to support long-term prosperity.

Many urgent issues in our world today, such as poverty, pollution, the environment, economic growth, standards of living, unemployment, inflation, technology, international trade and many more, involve economics.

The objective of economists is to explain, analyse and understand issues such as these in the hope of finding ways to deal with them so as to bring about improvements in the well-being of people everywhere.

Marrakesh, Morocco. Woolen handmade hats for sale in the market square of the old city

> Chapter 1

The foundations of economics

BEFORE YOU START

- As you begin this course, you may already have an idea of 'economics'. What do you think the subject is about?
- Sciences like physics, biology and chemistry are examples of 'natural sciences'. Anthropology, psychology and economics are examples of 'social sciences'. In what ways do you think natural and social sciences are similar and in what ways are they different?
- What do you think the purpose of government should be in society?

Chapter 1 of this book is an introduction to the social science of economics. We will discover the meaning of economics, and will discuss key concepts forming the basis of the economic perspective of the world. We will see how economists use models and theories to analyse economic problems, and will also learn about the organising principles of market, planned and mixed economies. The chapter will end with an account of famous economists who made important contributions to economic thought.

Photo: Bangkok, Thailand. Street foods collection November 2019

1.1 Understanding the nature of economics

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the social nature of economics (AO2)
- distinguish between microeconomics and macroeconomics (AO2)
- explain the nine central concepts that run through your course in economics (AO2)
- explain the meaning of scarcity (AO2)
    - as unlimited human needs and wants met by limited resources
    - in relation to sustainability
- identify and explain the four factors of production (AO2)
- explain the meaning of opportunity cost and its relationship to choice and free goods (AO2)

Economics as a social science

The social nature of economics

The **social sciences** are academic disciplines that study human society and social relationships. They are concerned with discovering general principles describing how societies function and are organised. The social sciences include anthropology, economics, political science, psychology, sociology and others.

Economics is a *social* science because it deals with human society and behaviour, and particularly those aspects concerned with how people organise their activities and how they behave to satisfy their needs and wants. It is a social *science* because its approach to studying human society is based on the scientific method, which we will consider below.

Microeconomics and macroeconomics

The study of economics breaks down the economic world into two levels. One of these is like looking at the economic world through a microscope, while the other is like looking at it through a telescope.

- The micro level, called **microeconomics**, examines the behaviour of individual decision-making units in the economy. The two main groups of decision-makers we study are consumers (or households) and firms (or businesses). Microeconomics is concerned with how these decision-makers behave, how they make choices, what are the consequences of their decisions and how their interactions in markets determine prices. (Micro comes from the Greek word *μικρό* or micró, meaning *small*.)
- The macro level, called **macroeconomics**, examines the economy as a whole to obtain a broad or overall picture of the economy. Macroeconomics uses *aggregates*, which are wholes or collections of many individual units, such as the sum of consumer behaviours and the sum of firm behaviours, and total income and output of the entire economy, as well as total employment and the overall price level. (Macro comes from the Greek word *μάκρος* or makros, meaning *large*.)

The ideas and principles that are developed in microeconomics and macroeconomics are the building blocks that economists use to study many specific areas of economics. Some of these areas are studied at the micro level, such as market failure (Chapters 5–7) and others at the macro level, such as unemployment and inflation (Chapter 10). In addition, the ideas and principles developed in microeconomics and macroeconomics are applied to the study of many other areas, such as International economics and Development economics, which you will discover in Unit 4 of this book.

Key concepts of this course

The concepts listed below will be discussed in a number of different contexts in your study of economics. While economists agree on the definitions of each of these, there are debates over how some of these concepts should be interpreted or applied, especially in connection with formulating appropriate economic policies to address important economic objectives. You will encounter many of these as you read this book.

Scarcity

One of the most important concepts in economics, **scarcity**, refers to the idea that resources are insufficient to satisfy unlimited human needs and wants. In fact, it is said that if there were no scarcity, there would be no social science of economics. This is because economics is the study of how our scarce or limited resources can best be used in order to satisfy the unlimited needs and wants of human beings.

Choice

In a very important sense, economics is the study of **choice**. Since resources are scarce, it is not possible for all human needs and wants to be satisfied. This means that choices must be made about what will be produced and what will be foregone (not produced and therefore sacrificed). Economics studies how different decision-makers make choices between competing alternative options, and analyses the present and future consequences of their choices.

Efficiency

Efficiency refers to making the best possible use of scarce resources to avoid resource waste. In view of the scarcity of resources, it is important to use these in ways that ensure they are not wasted. In part, efficiency means using the fewest possible resources to produce goods and services. But, in addition, it requires that scarce resources are used to produce the goods and services that mostly satisfy society's needs and wants. This is known as **allocative efficiency**, used as a benchmark or standard to determine the appropriateness of economic actions from the point of view of minimising resource waste.

Equity

Equity is the idea of being fair or just. Equity is not the same as equality, which is the sameness of treatment or outcomes for people or groups of people in a society. Fairness is a normative concept (to be discussed later) because ideas of what is fair vary according to beliefs, value judgements and ideologies. In economics, the ideas of equity and inequity are usually identified with equality and inequality, and are used mostly in connection with equality in the distribution of income, wealth and human opportunity. In all economic systems, these kinds of inequities or inequalities are present both within and between societies, and are significant issues, as many people cannot meet their basic needs and lack opportunities. There is much debate among economists on how much and what types of government intervention in markets are appropriate in order to address these issues effectively.

Economic well-being

Economic well-being is a concept that has several different dimensions. It refers to levels of prosperity, economic satisfaction and standards of living among the members of a society. Economic well-being includes:

- security with respect to income and wealth, having a job and housing
- the ability to pursue one's goals, work productively and develop one's potential
- the ability to have a satisfactory quality of life, which includes numerous factors such as health, education, social connections, environmental quality, personal security
- the ability to maintain all of the above over time.[1]

There are very significant variations in levels of economic well-being both within nations and between nations.

Sustainability

Sustainability refers to the long-term maintenance or viability of any particular activity or policy. In economics, it is most commonly used to refer to the ability of the present generation to satisfy its needs by the use of resources, and especially non-renewable resources, without limiting future generations' ability to satisfy their own needs. The problem arises because the present generation at any moment in time engages in many economic activities of production and consumption that too often destroy or degrade (lower the quality of) the environment and non-renewable resources. The result of such activities is that future generations will be penalised. Therefore the issue is how to develop methods of production and patterns of consumption that will not result in such environmental and resource destruction and degradation.

Change

'Panta rhei' is a famous saying by the Greek philosopher Heraclitus, that means 'everything flows'. Heraclitus taught that **change** is an essential part of life. This idea is very important in economics, where much of what we study is in a continuous state of change. In economics, we can distinguish between the idea of change: (i) in economic theory and (ii) in real-world

[1] OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth, Chapter 2 Economic well-being, OECD 2013.

events. In economic theory, economists very often study change between one situation and another situation that has been caused by a change in one or more variables. It is important to bear this in mind in your study of economics as you will often be asked to analyse and evaluate this kind of change in a large variety of contexts. Regarding the study of real-world phenomena, the world is characterised by continuous change in the institutional, technological, social, political and cultural environments in which economic events occur.

Interdependence

Interdependence refers to the idea that economic decision-makers interact with and depend on each other. Such interdependence occurs on all levels, from individuals, to communities, to nations and to groups of nations. Interdependence arises from the fact that no one is self-sufficient, requiring ever-increasing degrees of interactions and interdependence. Consumers, workers, firms, governments and all other individuals or groups of individuals depend on one another for the achievement of their economic goals. With increasing globalisation (which refers to the interactions and integration of economies world-wide), interdependence increases. In a highly interdependent world, events in one part give rise to many and possibly unintended consequences in other parts, with outcomes that cannot always be predicted or discovered by looking at the constituent parts in isolation. Economists must therefore take into consideration both intended and unintended consequences of economic decisions and events when there is a high degree of interdependence.

Intervention

In economics, **intervention** typically refers to government intervention, meaning that the government becomes involved with the workings of markets. While markets offer numerous advantages as a way to achieve important economic objectives, it is generally recognised that markets on their own often cannot achieve important societal goals, such as the goals of *equity*, *sustainability*, *economic well-being* or *efficiency*. When this occurs, whether at the micro or macro levels, there may be good reason for the government to intervene in order to correct for the market's deficiencies. However, economists and policy-makers often disagree widely on the need for, degree and method of intervention that is necessary. A key debate that you will repeatedly encounter in your studies of economics involves the advantages and disadvantages of free markets versus government intervention.

The fundamental problem of economics: scarcity and choice

The problem of scarcity

The term 'economics' is derived from the ancient Greek expression οίκον νέμειν (*oikon nemein*), which originally meant 'one who manages and administers all matters relating to a household'. Over time, this expression evolved to mean 'one who is prudent in the use of resources'. By extension, economics has come to refer to the careful management of society's scarce resources to avoid waste. Let's examine this idea more carefully.

Human beings have very many needs and wants. Some of these are satisfied by physical objects and others by non-physical activities. All the physical objects people need and want are called *goods* (food, clothing, houses, books, computers, cars, televisions, refrigerators and so on); the non-physical activities are called *services* (education, health care, entertainment, travel, banking, insurance and many more).

The study of economics arises because people's needs and wants are unlimited, or infinite. Whereas some individuals may be satisfied with the goods and services they have or can buy, most would prefer to have more: more and better computers, cars, educational services, transport services, housing, recreation, travel and so on; the list is endless.

Yet it is not possible for societies and the people within them to produce or buy all the things they want. Why is this so? It is because there are not enough **resources**. Resources are the inputs used to produce goods and services wanted by people, and for this reason are also known as **factors of production**. They include things like human labour, machines and factories, and 'gifts of nature' like agricultural land and metals inside the earth. Factors of production do not exist in unlimited abundance: they are *scarce*, or limited and insufficient in relation to unlimited uses that people have for them.

Scarcity is a very important concept in economics. It arises whenever there is not enough of something in relation to the need for it. For example, we could say that food is scarce in poor countries, or we could say that clean air is scarce in a polluted city. In economics, scarcity is especially important in describing a situation of *insufficient factors of production*, because this in turn leads to insufficient goods and services. Defining scarcity, we can therefore say that:

> **Scarcity** is the situation in which available resources, or factors of production, are finite, whereas wants are infinite. There are not enough resources to produce everything that human beings need and want.

Why scarcity forces choices to be made

The conflict between unlimited wants and scarce resources has an important consequence. Since people cannot have everything they want, they must make *choices*. The classic example of a choice forced on society by resource scarcity is that of 'guns or butter', or more realistically the choice between producing defence goods (guns, weapons, tanks) or food: more defence goods mean less food, while more food means fewer defence goods. Societies must choose how much of each they want to have. Note that if there were no resource scarcity, a choice would not be necessary, since society could produce as much of each as was desired. But resource scarcity forces the society to make a choice between available alternatives. Economics is therefore a study of choices.

The conflict between unlimited needs and wants, and scarce resources has a second important consequence. Since resources are scarce, it is important to avoid waste in how they are used. If resources are not used effectively and are wasted, they will end up producing less; or they may end up producing goods and services that people do not really want or need. Economics must try to find how best to use scarce resources so that waste can be avoided. Defining economics, we can therefore say that:

> **Economics** is the study of choices leading to the best possible use of scarce resources in order to best satisfy unlimited human needs and wants.

As you can see from this definition of economics, economists study the world from a social perspective, with the objective of determining what is in society's best interests.

TEST YOUR UNDERSTANDING 1.1

1. Think of some of your most important needs and wants, and then explain whether these are satisfied by goods or by services.
2. Outline why economics is a study of choices. Describe its relationship to scarcity. Outline how scarcity is related to the need to avoid waste in the use of resources.
3. Explain why diamonds are far more expensive than water, even though diamonds are a luxury while water is a necessity without which we cannot live.

Scarcity and sustainability

The meaning of sustainability

Economic activities in many (if not most) countries are often achieved at the expense of the natural environment and natural resources. *Economic growth*, which involves increases in the amount of goods and services produced, very often results in increased air and water pollution, and the destruction or depletion of forests, wildlife and the ozone layer, among many other natural resources. Increasing awareness of this issue has given rise to the concept of *sustainable development*, defined as 'development which meets the needs of the present without compromising the ability of future generations to meet their own needs'.[2]

Sustainable development occurs when societies grow and develop without leaving behind fewer or lower-quality resources for future generations. If we, in the present, use up resources at a rate that leaves fewer or lower-quality resources behind, we are satisfying our needs and wants now at the expense of people in the future, who with fewer or lower-quality resources will be less able to satisfy their own needs and wants. If we change the global climate and use up clean air, seas and rivers, forests and the ozone layer, we put future generations at a disadvantage and even in danger.

[2] Brundtland Commission (World Commission on Environment and Development) (1987) *Our Common Future*, Oxford University Press

Using the definition of sustainable development, we can see that *sustainability* or *sustainable resource use* involves using resources in ways and at rates that do not reduce their quantity or quality over time. As a rule, this term is used with reference to renewable resources, or those kinds of natural resources that are able to reproduce themselves (such as forests, fish and sea life, air quality, the fertility of the soil). Sustainable resource use does not mean that these kinds of natural resources should not be used at all, but rather that they should be used at a rate that gives them enough time to reproduce themselves, so that they can be maintained over time in terms of quantity and quality and not be destroyed or depleted.

It is clear that the issue of sustainable use of resources arises from the fact that these resources are *scarce*. If they were not scarce, it would not matter at all how fast we used them or destroyed them as there would be plenty more.

Sustainability refers to maintaining the ability of the environment and the economy to continue to produce and satisfy needs and wants into the future; sustainability depends crucially on **sustainable resource use**, referring to the preservation of the environment over time. The problem of sustainability arises because resources are scarce.

Threats to sustainability do not only result from high-income production and consumption patterns that rely strongly on polluting fossil fuels as well as other activities that destroy the environment. In addition, in very poor societies, threats to sustainability often arise from poverty itself, which drives very poor people to destroy their natural environment as they make efforts to survive. Examples include cutting down forests, overgrazing, soil erosion and many more. In all these cases, there may be an unsustainable use of resources, as fewer and lower-quality resources are left behind for future generations.

While virtually everyone today agrees on the importance of sustainability, there is vast disagreement about what this means from a practical point of view, and how this can be achieved in practice. We will discuss the issue of sustainable resource use and policies to achieve this in Chapter 5.

TEST YOUR UNDERSTANDING 1.2

1 Explain the relationship between scarcity and sustainability.

2 Consider the following: 'Dangerous levels of industrial air pollution in India reduce life expectancy by seven years for 40% of its population. This refers to the people living in the states of the Indo-Gangetic plain, where pollution has increased by 72% in a period of 18 years. The population of Delhi faces an average loss of 10.2 years. The public is outraged. The government has begun to respond to this public health emergency'.[3]

   a Research an example of a challenge India faces and outline how it is a threat to sustainable development.

   b Describe how India's rapid economic growth impacts its future generations.

Resources as factors of production

We have seen that resources, or all inputs used to produce goods and services, are also known as factors of production.

The four factors of production

Economists group the factors of production under four broad categories:

- **Land** consists of all natural resources, including all agricultural and non-agricultural land, as well as everything that is under or above the land, such as minerals, oil reserves, underground water, forests, rivers and lakes. Natural resources are also called 'gifts of nature'.
- **Labour** includes the physical and mental effort that people contribute to the production of goods and services. The efforts of a teacher, a construction worker, an economist, a doctor, a taxi driver or a plumber all contribute to producing goods and services, and are all examples of labour.

[3] https://www.indiatoday.in/india/story/pollution-from-punjab-to-bengal-48-crore-people-may-die-7-years-early-but-all-is-not-lost-study-1614689-2019-11-01
https://ig.ft.com/india-pollution

- **Capital**, also known as *physical capital*, is a man-made factor of production (it is itself produced) used to produce goods and services. Examples of physical capital include machinery, tools, factories, buildings, road systems, airports, harbours, electricity generators and telephone supply lines. Physical capital is also referred to as a capital good or investment good.
- **Entrepreneurship** (management) is a special human skill possessed by some people, involving the ability to innovate by developing new ways of doing things, to take business risks and to seek new opportunities for opening and running a business. Entrepreneurship organises the other three factors of production and takes on the risks of success or failure of a business.

Other meanings of the term 'capital'

The term 'capital', in the most general sense, refers to resources that can produce a future stream of benefits. Thinking of capital along these lines, we can understand why this term has a variety of different uses, which although are seemingly unrelated, in fact all stem from this basic meaning.

- **Physical capital**, defined above, is one of the four factors of production consisting of man-made inputs that provide a stream of future benefits in the form of the ability to produce greater quantities of output: physical capital is used to produce more goods and services in the future.
- **Human capital** refers to the skills, abilities and knowledge acquired by people, as well as good levels of health, all of which make them more productive. Human capital provides a stream of future benefits because it increases the amount of output that can be produced in the future by people who embody skills, education and good health.
- **Natural capital**, also known as *environmental capital*, refers to an expanded meaning of the factor of production 'land' (defined earlier). It includes everything that is included in land, plus additional natural resources that occur naturally in the environment such as the air, biodiversity, soil quality, the ozone layer and the global climate. Natural capital provides a stream of future benefits because it is necessary to humankind's ability to live, survive and produce in the future.
- **Financial capital** refers to investments in financial instruments, like stocks and bonds, or the funds (money) that are used to buy financial instruments. Financial capital also provides a stream of future benefits, which take the form of an income for the holders, or owners, of the financial instruments.

TEST YOUR UNDERSTANDING 1.3

1. a Outline why resources are also called 'factors of production'.
   b Identify the four factors of production.
2. Outline how physical capital differs from the other three factors of production.
3. Describe why entrepreneurship is considered to be a factor of production separate from labour.
4. a Identify the various meanings of the term 'capital'.
   b Outline what they have in common.

Scarcity, choice and opportunity cost: the economic perspective

Opportunity cost

Opportunity cost is defined as the value of the next best alternative that must be given up or sacrificed in order to obtain something else. Every time we choose to do something, we give up something else we could have done instead. For example, your decision to read this book now means you have given up a different activity, such as sleeping, watching TV or visiting a friend. If your best or favourite alternative to reading this book is watching TV, the TV time you have sacrificed is the opportunity cost of reading this book. Opportunity cost in this case rises from the fact that time is limited or scarce; if it were endless, you would never have to sacrifice any activity in order to do something else.

When a consumer chooses to use her $100 to buy a pair of shoes, she is also choosing not to use this money to buy books, food or anything else; if books are her favourite alternative to shoes, the books she sacrificed (did not buy) are the opportunity cost of the shoes. Note that if the consumer had endless amounts of money, she could buy everything she wanted and the shoes would have no opportunity cost.

The concept of **opportunity cost**, or the value of the next best alternative that must be sacrificed to obtain something else, is central to the economic perspective of the world, and results from the scarcity that forces choices to be made.

TEST YOUR UNDERSTANDING 1.4

1. Define opportunity cost.
2. Explain how scarcity and choice relate to opportunity cost.
3. Think of three choices you have made during the past week and describe the opportunity cost of each one.

Free and economic goods

Based on the concept of opportunity cost, we can make a distinction between free goods and economic goods (note that the term 'good' is used here in a general sense to include goods, services and resources):

A **free good** is any good that is not scarce, and therefore has a zero opportunity cost. Since it is not limited by scarcity, it includes anything that can be obtained without sacrificing something else. An *economic good* is any good that is scarce, either because it is a naturally-occurring scarce resource (such as oil, gold, coal, forests, lakes), or because it is produced by scarce resources. All economic goods have an opportunity cost greater than zero.

Free goods are rare. Sometimes a good can be a free good in certain situations and an economic good in others. For example, arable land in America before European colonisers arrived was a free good because it was so abundant; as the colonisers grew in numbers it became increasingly scarce and therefore an economic good. Salt used to be a free good that has become an economic good. Oxygen in the open unpolluted countryside can be a free good; in a room with no windows that is crowded with people, it becomes an economic good. Unobstructed sunshine is also a free good in many situations.

It is important to distinguish free goods from goods or resources that are available free of charge to their users. There are two categories of goods that are available free of charge, but which do have opportunity costs and are therefore economic goods:

- goods provided by the government, such as the road system, public parks and playgrounds, free education, free health care services; all these are economic goods produced by scarce resources, and are paid for out of tax revenues (see Chapter 6);
- certain natural resources, such as clean air, forests, rivers, lakes and wildlife, that are not owned by anyone (they are called *common pool resources*; these are also economic goods because they are scarce, and are becoming increasingly scarce due to overuse and depletion (see Chapter 5).

TEST YOUR UNDERSTANDING 1.5

1. Explain the difference between a free good and an economic good.
2. Identify which of the following goods are 'free goods' and explain why:
   a. public parks
   b. sand in the Sahara desert
   c. garbage collection
   d. free health care services
   e. wildlife.
3. Why do you think free goods are rare?

1.2 The three basic economic questions: resource allocation and output/income distribution

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- identify and explain the three basic economic questions (AO2)
- distinguish between the role of markets and government intervention in designing and proposing solutions to the basic economic questions (AO2)
- distinguish between economic systems: the free market economy, planned economy and mixed economy (AO2)

The basic economic questions: what/how much, how and for whom to produce

Scarcity forces every economy in the world, regardless of its form of organisation, to answer the following three basic questions:

- **What/how much to produce.** All economies must choose what particular goods and services and what quantities of these they wish to produce.
- **How to produce.** All economies must make choices on how to use their resources in order to produce goods and services. Goods and services can be produced by use of different combinations of factors of production (for example, relatively more labour with fewer machines, or relatively more machines with less labour), by using different skill levels of labour, by using different technologies or by using different raw materials, for example, plastic or wood.
- **For whom to produce.** All economies must make choices about how the goods and services produced are to be distributed among the population. Should everyone get an equal amount of these? Should some people get more than others? Should some goods and services (such as education and health care services) be distributed more equally?

Resource allocation and output/income distribution

The first two of these questions, *what/how much to produce* and *how to produce*, are about *resource allocation*, while the third, *for whom to produce*, is about the *distribution of output and income*.

Resource allocation refers to assigning available resources, or factors of production, to specific uses chosen among many possible alternatives, and involves answering the *what/how much to produce* and *how to produce* questions. For example, if a *what/how much to produce* choice involves choosing a certain amount of food and a certain amount of weapons, this means a decision is made to *allocate* some resources to the production of food and some to the production of weapons. At the same time, a choice must be made about *how to produce*: which particular factors of production and in what quantities (for example, how much labour, how many machines, what types of machines, etc.) should be assigned to produce food, and which and how many to produce weapons.

If a decision is made to change the amounts of goods produced, such as more food and fewer weapons, this involves a *reallocation* of resources. Sometimes, societies produce the ‘wrong’ amounts of goods and services relative to what is socially desirable. For example, if too many weapons are being produced, we say there is an *overallocation* of resources to production of weapons. If too few socially desirable goods or services are being produced, such as education or health care, we say there is an *underallocation* of resources to the production of these.

The third basic economic question, *for whom to produce*, involves the *distribution of output* and is concerned with how much output different individuals or different groups in the population receive. This question is also concerned with the **distribution of income** among individuals and groups in a population, since the amount of output people can get depends on how much of it they can buy, which in turn depends on the amount of income they have. When the distribution of income or output changes so that different social groups now receive more, or less, income and output than previously, this is referred to as **redistribution of income**.

TEST YOUR UNDERSTANDING 1.6

1. State the three basic economic questions that must be addressed by any economy.
2. Explain the relationship between the three basic economic questions, the allocation of resources and the distribution of income or output.
3. Consider the following, and identify each one as referring to output/income distribution or redistribution; or to resource allocation, reallocation, overallocation or underallocation (note that there may be more than one answer).
   - a Evidence suggests that over the last two decades in many countries around the world the rich are getting richer and the poor are getting poorer.
   - b In Brazil, the richest 10% of the population receive 48% of total income.
   - c Whereas rich countries typically spend 8–12% of their income on providing health care services to their populations, many poor countries spend as little as 2–3% of their income.
   - d Many developing countries devote a large proportion of their government budget funds to spending on university level education, while large parts of their population remain illiterate.
   - e If countries around the world spent less on defence, they would be in a position to expand provision of social services, including health care and education.
   - f Pharmaceutical companies spend most of their research funds on developing medicines to treat diseases common in rich countries, while ignoring the treatment of diseases common in poor countries.

Alternative ways to answer the economic questions

Countries around the world differ enormously in the ways they make allocation and distribution decisions. At the heart of their differences lie the methods used to make the choices required by the *what/how much*, *how* and *for whom to produce* questions. There are two main methods that can be used to make these choices: the *market method* and the *command method*.

Markets versus government intervention

The meaning of government intervention in the market

In the market method, resources are owned by private individuals or groups of individuals, and it is mainly consumers and firms (or businesses) who make economic decisions by responding to prices that are determined in markets (we will see how this happens in Chapter 2). In the command method, resources (land and capital in particular) are owned by the government, which makes economic decisions by commands. In practice, commands involve legislation and regulations by the government, or in general any kind of government decision-making that affects the economy.

In the real world, there has never been an economy that is entirely a market economy or entirely a command economy. Real-world economies combine markets and commands in many different ways, and each country is unique in the ways they combine them. Economies may lean more toward the command economy (as in planned economies of communist systems), or more toward the market economy (as in highly market-oriented economies). Whatever the case, in the last 40 or so years, there has been a trend around the world for economies to rely more and more on the market and less on commands.

When the government makes decisions that affect the economy, this is known as *government intervention*, because the government intervenes (or interferes) in the workings of markets. Examples of government intervention include provision of public education, public health care, public parks, road systems, national defence, flood control, minimum wage legislation, restrictions on imports, anti-monopoly legislation, tax collection, income redistribution and many more.

Whatever the reasons for and types of government intervention in the market, **government intervention** changes the allocation of resources (and distribution of output and income) from what markets would have achieved working on their own.

The market economy offers important benefits that we will discover in Chapter 2. Yet it does not always produce the 'best' answers to the *what/how much*, *how* and *for whom* questions for many reasons to be discussed in later chapters. Therefore, a market economy cannot operate effectively without some government intervention.

Whereas practically everyone agrees that some government intervention in markets is necessary, economists disagree widely over how much governments should intervene and how they should intervene. There are two broad schools of thought on this issue. One focuses on the positive aspects of markets, while the other focuses on the imperfections of markets.

According to the first perspective, economists argue that in spite of imperfections, markets are able to work reasonably well on their own, and can produce outcomes that generally promote society's well-being. Markets can achieve a reasonably good allocation of resources, answering the *what/how much to produce* and *how to produce* questions quite well. Government intervention changes this allocation of resources, and often worsens it, giving rise to resource waste. Therefore, while some minimum government intervention may be needed in certain situations, this should not be very extensive.

According to the second school of thought, markets have the potential to work well, but in the real world their imperfections may be so important that they make government intervention necessary for their correction. This means that markets, working on their own, do not do a very good job of allocating resources in society's best interests. The purpose of government intervention, therefore, is to help markets work better and arrive at a better pattern of resource allocation and distribution of income and output.

Economic systems: free market economy, planned economy, mixed economy

It is suggested that you reread this section after reading Chapters 2 and 4–7, as you will then be better able to understand it.

The market and command methods to answer the basic economic questions discussed above can be thought of as two 'ideal types' of economies: a **free market economy** based on the market approach and a **planned economy** based on the command approach. An ideal type is an abstract idea that does not claim to represent the real world, but rather contains some characteristics that serve as a standard for comparison of real-world situations. (Note that an ideal type is not 'ideal' in the sense of perfect or excellent.) As we also discussed above, real-world economies combine markets and commands in many different ways, resulting in **mixed economies**.

The ideal-type free market and planned economies are distinguished from each other on the basis of three criteria.

- ***Resource ownership: public or private sector***
  Ownership of society's resources can be 'public' or 'private'. The *public sector* refers to the parts of the economy that are under the ownership of the government (whether national, regional or local). The government is also sometimes referred to as the 'state'. The *private sector* includes the parts of the economy that are under the ownership of private individuals or groups of individuals; these include consumers (households), firms (businesses) and resource owners, as well as organisations such as NGOs (non-governmental organisations) and interest groups (for example, consumer protection organisations). *The free market economy has private sector resource ownership, and the planned economy has public sector resource ownership.*

TEST YOUR UNDERSTANDING 1.7

1 Identify some more examples of command methods (government intervention) in market economies.

2 Outline the main source of the disagreement between those who argue there should be little government intervention in the economy and those who argue that government intervention should be more extensive.

- ***Economic decision-making***
  Economic decisions and choices regarding the *what/how much, how* and *for whom* questions can be made by the public sector, i.e. the government, or by the private sector. There are many private decision-makers, as noted above, but the most important of these are consumers (households), firms (businesses) and resource owners. *The free market economy has private sector economic decision-making and the planned economy has public sector economic decision-making*.
- ***Rationing systems***
  The term **rationing**[4] can be defined as a method used to apportion or divide something up between its interested users. In economics, it refers to the method used to make resource allocation and output/income distribution decisions. There are two kinds of *rationing systems*: price rationing and non-price rationing.
  - The *free market economy* uses price rationing to make resource allocation and output/income distribution decisions. This means that all economic decisions relating to what will be produced, how it will be produced, and who will receive the output are made on the basis of prices of goods, services and resources that have been determined in markets.
  - The *planned economy* uses non-price rationing to make resource allocation and output/income distribution decisions. This means that all decisions relating to what/how much will be produced, how it will be produced, and who will receive the output, are made by use of methods that have nothing to do with prices determined in markets. Non-price rationing results when there are no markets, or when governments interfere in markets, in which case the government acts as a central authority and makes economic decisions by commands.

Both price rationing and non-price rationing will become much clearer to you after you have studied Chapter 2.

In the *free market economy* households and firms (the private sector) are the main owners of resources, as well as the economic decision-makers who make buying and selling decisions, and who are linked together in product and resource markets. As we will see in Chapter 2 , product and resource markets determine prices of goods, services and resources, which act as the basis of price rationing.

The *planned economy* is characterised by the absence of markets or the limited operation of markets. As the owner of resources and economic decision-maker, the government makes all allocation and distribution decisions through non-price rationing. This is called a planned economy because the government authority makes detailed plans of all economic activities, on the basis of which it directs and coordinates economic decisions through commands. There remain very few countries in the world that are still highly centrally planned; these include North Korea and, to a lesser extent, Cuba. Most other countries have begun to introduce major reforms intended to strengthen the role of markets.

In the real world, virtually all economies combine elements of both markets and commands. Differences between actual economies lie mainly in the ways the two are combined and in the degree to which one predominates. For this reason, most economies in the world today are *mixed economies*.

The mixed economy and mixed market economy

Increasingly, mixed economies are becoming *mixed market economies*. That is, most economic activity is market-based based rather than centralised, and price rationing is more common than non-price rationing.

In mixed market economies, public or private sector ownership and decision-making often go together. For example, privately owned firms usually make decisions about what and how they will produce and sell, while the government makes decisions about economic activities that fall under its ownership (such as public health services, public road systems, public parks, defence facilities and many others).

However, the government's decision-making role in the mixed market economy is not limited to activities falling under its ownership; it also extends into private sector activities. For example, the United States, one of the more market-oriented countries in the world, has government decision-making that affects the private sector in numerous areas such as minimum wage legislation, subsidies for agricultural products, tariffs on imports, regulation of private sector activities, anti-monopoly legislation, tax collection, income redistribution and many others. All mixed market economies, in fact, have government involvement with the private sector that is either a response to the failure of the market mechanism to work well, or in response to the demands of politically powerful interest groups. We will examine government intervention in the market extensively for both these sets of reasons in later chapters.

[4] Will be introduced in Chapter 2 as a syllabus term.

Government involvement in the private sector varies widely from country to country in extensiveness. For example, the free market plays a more prominent role in the United Kingdom and the United States compared to France and Japan. Also, government involvement in the private sector varies in the form that it takes. For example, in the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden), there is extensive government intervention in income redistribution; in Japan, extensive government intervention takes the form of planning and coordinating private sector activities.

In mixed market economies, both price and non-price rationing can be observed, but with price rationing predominating. In general, price rationing arises in situations where there is a market for resources, goods and services. If there is no market (or if markets are not free because of government intervention), then some form of non-price rationing occurs. For example, when governments in market economies provide national defence, public health care systems, public road systems and flood control, they do not rely on price rationing to determine resource allocation and output distribution, and the role of the government agencies that plan and provide these services is similar to the role of the central planner. Consider the case of national health systems, where the government, through tax financing, undertakes to provide health care services that are made available to the entire population free (or nearly free) of charge. Since there is no price charged to the consumer who receives a service, some mechanism other than price, i.e. non-price rationing, must be used to distribute the service among its users. The most commonly used non-price mechanism is the waiting line or waiting period (queue).

There have been significant changes over time in the relative prominence of private versus public sector

TEST YOUR UNDERSTANDING 1.8

1. Identify the basic economic questions that must be answered by
   a. free market economies,
   b. command economies, and
   c. mixed economies.
2. Use the criteria appearing in Table 1.1 to compare and contrast the main characteristics of the market and command economies.
3. Outline why the command economy is also referred to as a centrally planned economy.
4. Compare and contrast the methods by which allocation and distribution choices are made in the market economy and the command economy.
5. a. Define and explain the difference between price rationing and non-price rationing.
   b. Describe the functions of these two rationing mechanisms.
   c. Identify the kind of economic system in which each one predominates.
6. Describe examples of command practices in mixed economies.
7. Outline why most economies in the world today are mixed market economies.

Criteria	Free market economy (based on prices)	Planned economy (based on commands)	Mixed economy and mixed market economy (based on prices and commands)
Resource ownership	private sector	public sector	public and private sectors
Economic decision-making	private sector	public sector	public and private sectors
Rationing system	price rationing	non-price rationing	price rationing and non-price rationing

Table 1.1: Free market, planned and mixed economies

activities. During much of the 20th century, many countries throughout the world saw major increases in government participation in economic decision-making. Since the 1980s, there has been a shift once again in the direction of less public sector involvement and a corresponding growth in private sector activities. In many countries around the world, including both more developed and less developed ones, the increasing importance of market-based activities has been due to the growing popularity of many supply-side economic policies (to be discussed in Chapter 13), as well as a recognition of the limitations of central planning. Table 1.1 provides a summary of the three criteria as they apply to each type of economy.

Evaluating the free market economy and planned economy (Supplementary material)

The interested student may explore this topic in the digital coursebook.

1.3 Understanding the world by use of models

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- identify and explain the relationships illustrated in the production possibilities curve model PPC (AO2)
- use the PPC model to explain opportunity cost, scarcity, choice, unemployment, efficiency, actual growth and growth in production possibilities (AO2)
- draw a diagram to explain all the above concepts in the PPC model (AO4)
- distinguish between increasing versus constant opportunity cost in the PPC model (AO2)
- draw a diagram to illustrate the difference between increasing and constant opportunity cost in the PPC model (AO4)

CONTINUED

- identify and explain the interdependent activities of decision-makers in the circular flow of income model: households, firms, the government, banks and the financial sector, and the foreign sector (AO2)
- explain the role of leakages and injections in the circular flow of income model (AO2)
- draw a diagram to identify and explain the relationships illustrated in the circular flow of income model including leakages and injections (AO4)

Everyone is familiar with the idea of a model. As children, many of us played with paper aeroplanes, which are models of real aeroplanes. In chemistry at school, we studied molecules and atoms, which are models of what matter is made of. A **model** is a simplified representation of something in the real world; it represents only the important aspects of the real world being investigated, ignoring unnecessary details. This way it allows us to focus on important relationships. Models are used a lot by scientists and social scientists in their efforts to understand or explain real-world situations.

Introducing the production possibilities curve model

Consider a simple hypothetical economy producing only two goods: microwave ovens and computers. This economy has a fixed (unchanging) quantity and quality of resources (factors of production) and a fixed technology (the method of production is unchanging). Table 1.2 shows the combinations of the two goods this economy can produce. Figure 1.1 plots the data of Table 1.2: the quantity of microwave ovens is plotted on the vertical axis, and the quantity of computers on the horizontal axis.

If all the economy's resources are used to produce microwave ovens, the economy will produce 40 microwave ovens and 0 computers, shown by point A. If all resources are used to produce computers, the economy will produce 33 computers and 0 microwave ovens; this is point E. All the points on the curve joining A and E represent other production possibilities where some of the resources are used to produce microwave ovens and the rest to produce computers.

Point	Microwave ovens	Computers
A	40	0
B	35	17
C	26	25
D	15	31
E	0	33

Table 1.2: Combinations of microwave ovens and computers

For example, at point B there would be production of 35 microwave ovens and 17 computers; at point C, 26 microwave ovens and 25 computers, and so on. The line joining points A and E is known as the *production possibilities curve* (*PPC*) (or *production possibilities frontier*, *PPF*).

In order for the economy to produce the greatest possible output, in other words somewhere on the *PPC*, two conditions must be met:

- **All resources must be fully employed.** This means that all resources are being fully used. If there were unemployment of some resources, in which case they would be sitting unused, the economy would not be producing the maximum it can produce.
- **All resources must be used efficiently.** Specifically, there must be efficient resource use. The term 'efficiency' in a general sense means that resources are being used in the best possible way to avoid waste. (If they are not used in the best possible way, we say there is 'inefficiency'.) Efficiency in production means that output is produced by use of the fewest possible resources; alternatively, we can say that output is produced at the lowest possible cost. If output were not produced using the fewest possible resources, the economy would be 'wasting' some resources.

The **production possibilities curve (or frontier)** represents all combinations of the maximum amounts of two goods that can be produced by an economy, given its resources and technology, when there is full employment of resources and efficiency in production. All points on the curve are known as **production possibilities**.

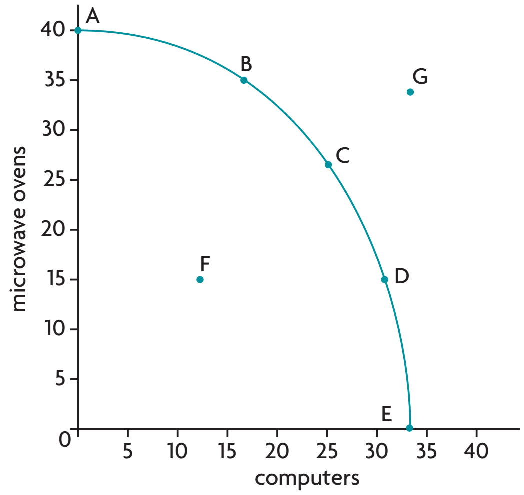


Figure 1.1: Production possibilities curve

What would happen if either of the two conditions (full employment and efficiency) is not met? Very simply, the economy will not produce at a point on the *PPC*; it will be somewhere inside the *PPC*, such as at point F. At F, the economy is producing only 15 microwave ovens and 12 computers, indicating that there is either unemployment of resources, or inefficiency in production, or both. If this economy could use its resources fully and efficiently, it could, for example, move to point C and produce 26 microwave ovens and 25 computers.

However, in the real world no economy is ever likely to produce on its *PPC*.

An economy's *actual output*, or the quantity of output actually produced, is always at a point inside the *PPC*, because in the real world all economies have some unemployment of resources and some inefficiency in production. The greater the unemployment or the inefficiency, the further away is the point of production from the *PPC*.

The production possibilities curve and scarcity, choice and opportunity cost

The production possibilities model is very useful for illustrating the concepts of scarcity, choice and opportunity cost:

- **Because of scarcity, the economy cannot produce outside its PPC.** With its fixed quantity and quality of resources and technology, the economy cannot move to any point outside the *PPC*, such as point G on Figure 1.1, because it does not have enough resources (i.e. there is resource scarcity).

- **Because of scarcity, the economy must make a choice about what particular combination of goods will be produced.** Assuming it could achieve full employment and efficiency, it must decide at which particular point on the *PPC* it should produce. (In the real world, the choice would involve a point inside the *PPC*.)
- **Because of scarcity, choices involve opportunity costs.** If the economy were at any point on the curve, it would be impossible to increase the quantity produced of one good without decreasing the quantity produced of the other good. In other words, when an economy increases its production of one good, there must be a sacrifice of some quantity of the other good. This sacrifice is the opportunity cost.

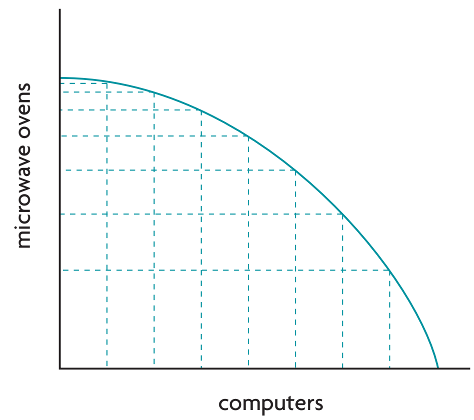
Let's consider the last point more carefully. Say the economy is at point C, producing 26 microwave ovens and 25 computers. Suppose now that consumers would like to have more computers. It is impossible to produce more computers without sacrificing production of some microwave ovens. For example, a choice to produce 31 computers (a move from C to D) involves a decrease in microwave oven production from 26 to 15 units, or a sacrifice of 11 microwave ovens. The sacrifice of 11 microwave ovens is the opportunity cost of 6 extra computers (increasing the number of computers from 25 to 31). Note that opportunity cost arises when the economy is on the *PPC* (or more realistically, somewhere close to the *PPC*). If the economy is at a point inside the curve, it can increase production of both goods with no sacrifice, hence no opportunity cost, simply by making better use of its resources: reducing unemployment or increasing efficiency in production.

The shape of the production possibilities curve

In Figure 1.2(a) the *PPC*'s shape is similar to that of Figure 1.1, while in Figure 1.2(b) it is a straight line. When the *PPC* bends outward and to the right, as in Figure 1.2(a), opportunity costs change as the economy moves from one point on the *PPC* to another. In part (a), for each additional unit of computers that is produced, the opportunity cost, consisting of microwave ovens sacrificed, gets larger and larger as computer production increases. This happens because of specialisation of factors of production, which makes them not equally suitable for the production of different goods and services.

As production switches from microwave ovens to more computers, it is necessary to give up increasingly more microwave ovens for each extra unit of computers produced, because factors of production suited to microwave oven production will be less suited to computer production. By contrast, when the *PPC* is a straight line (as in Figure 1.2(b)), opportunity costs are constant (do not change) as the economy moves from one point of the *PPC* to another. Constant opportunity costs arise when the factors of production are equally well suited to the production of both goods, such as in the case of basketballs and volleyballs, which are very similar to each other, therefore needing similarly specialised factors of production to produce them. As we can see in Figure 1.2(b), for each additional unit of volleyballs produced, the opportunity cost, or sacrifice of basketballs, does not change.

a Increasing opportunity costs



b Constant opportunity costs

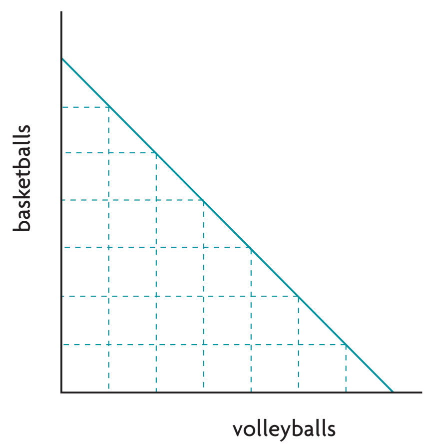


Figure 1.2: Production possibilities curve with increasing and constant opportunity costs

Explaining economic growth by use of the production possibilities curve model

Economic growth refers to increases in the quantity of output produced in an economy over a period of time. What are the causes of growth? We can find the answer to this question in the production possibilities model.

We have seen that any economy is most likely to be actually situated at some point inside its *PPC*, as it is very difficult for an economy ever to be fully efficient and have maximum employment of all resources. The further away an economy is situated from its *PPC*, the greater its resource unemployment and inefficiency. Therefore, by reducing unemployment and increasing efficiency, a country moves closer to its *PPC* and increases the actual quantity of output produced. It follows that *reductions in unemployment and increases in efficiency are two factors that can cause growth of actual output*. In Figure 1.3(a), the movement from point A to point B illustrates *actual growth*.

However, reduction of unemployment and inefficiencies can only result in a limited amount of economic growth. As the economy moves closer to its *PPC*, the ability to achieve more growth is exhausted, and more growth can only occur if there is *growth in production possibilities*, illustrated by an outward shift of the *PPC*. An outward shift of the *PPC* means the economy can produce more of both goods (*X* and *Y*), shown in Figure 1.3(b) by the shifts from PPC_1 to PPC_2 to PPC_3. In this figure, the growth in production possibilities is accompanied by outward movements of the economy's points of actual production, from A to B to C.

The factors that lead to outward shifts of the *PPC*, or growth in production possibilities are:

- increases in the quantity of resources (factors of production) in the economy
- improvements in the quality of resources (for example, through more educated labour)
- technological improvements.

As its production possibilities grow, unemployment must be kept at low levels and inefficiencies should be reduced to ensure that actual output continues to grow along with production possibilities, as in Figure 1.3(b). For example, an increase in the size of the labour force will do little to increase actual output produced if much of this labour remains unemployed (in this case the economy could remain stuck at point A even as PPC_1 shifts to PPC_2). Similarly, the discovery of major oil

> It is important to distinguish between *actual growth*, which involves a movement from one point inside the PPC to another point closer to the PPC, and *growth in production possibilities* involving an outward shift of the PPC. **Actual growth** is caused by reduction in unemployment and increases in efficiency in production. **Growth in production possibilities** is caused by increases in the quantity of resources, improvements in the quality of resources and technological improvements.

a Economic growth as an increase in actual output caused by reductions in unemployment and inefficiency in production

b Economic growth as an increase in production possibilities caused by increases in resource quantities or improvements in resource quality or technological improvements

c Decrease in production possibilities

d Non-parallel shifts of the *PPC*

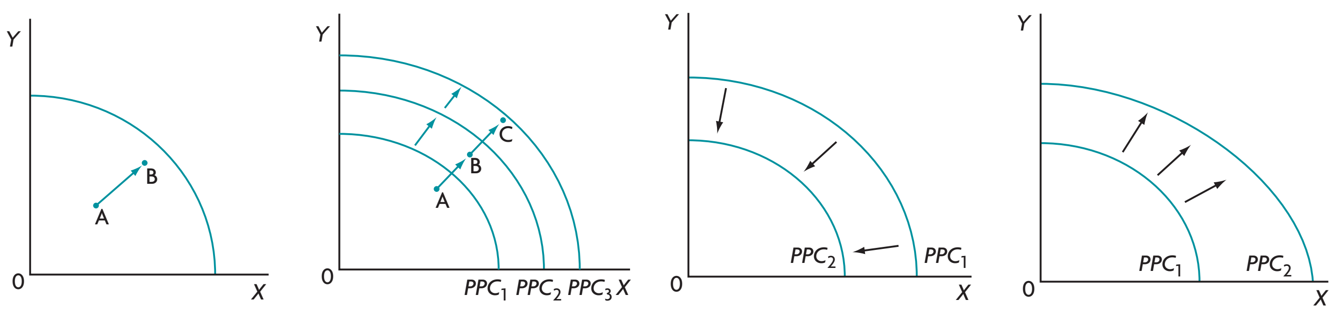


Figure 1.3: Using the production possibilities model to illustrate economic growth

reserves may do little to expand actual output if these reserves remain unexploited, or if their exploitation is undertaken inefficiently.

The *PPC* can also shift inward, indicating a decrease in production possibilities, or that less of the two goods can be produced, as shown in Figure 1.3(c). This results from a decrease in the quantity of resources or deterioration in resource quality.

An outward or inward shift need not be parallel; this is illustrated in Figure 1.3(d). For example, a technological change favouring the production of one good (*X*) increases the production of that good proportionately more. Similarly, an influx of unskilled workers into a country results in a larger proportionate increase in the production of goods using relatively more unskilled labour.

TEST YOUR UNDERSTANDING 1.9

1 Consider the production possibilities data in Table 1.2 and Figure 1.1. If the economy is initially at point A and moves to point B, computer production will increase by 17 units.

   a Calculate the opportunity cost of the increase in computer production.

   b If the economy moves from D to C, calculate the gain and its opportunity cost.

   c If it moves from point C to B, calculate the gain and its opportunity cost.

2 Use the concept of opportunity cost to explain why the following two statements have the same meaning:

   a efficiency in production means producing by use of the fewest possible resources

   b efficiency in production means producing at the lowest possible cost.

3 a Distinguish between output actually produced and output on the *PPC*.

   b Outline why an economy's actual output is most likely to be located somewhere inside its *PPC*.

4 Say an economy is initially at point F, producing 15 microwave ovens and 12 computers (Figure 1.1). State what would be the opportunity cost of moving to a point on the production possibilities curve, such as point C, where it would be producing 26 microwave ovens and 25 computers.

5 a Using diagram(s), distinguish between actual growth and growth in production possibilities.

   b List the factors that can give rise to each of these.

6 Use the production possibilities curve model and diagrams to show how the following can result in actual growth or growth (or decrease) in production possibilities:

   a a discovery of new oil reserves

   b firms hire more workers

   c a vaccine for contagious diseases is invented

   d firms improve how they manufacture and lower their costs of production

   e the widespread use of a new technology

   f a violent conflict destroys a portion of a country's factories, machines and road system

   g large cuts in government spending on education and health care lower levels of education and health in a population

   h an increase in the quantity of capital goods

   i an improvement in the level of education and skills of workers

   j industrial pollution destroys the environment.

7 Using diagrams, distinguish between increasing and constant opportunity costs.

8 Using the production possibilities model, explain the relationship between scarcity, choice and opportunity cost.

The circular flow of income model

The **circular flow of income model** is a simple model that illustrates some economic concepts and relationships that will help us understand the overall economy. In its simplest version, shown in Figure 1.4, the model illustrates a *closed economy*, meaning it has no links with other countries (it is 'closed' to international trade), and is also a model of an economy with no government and no banks or financial market.

It is assumed that the only decision-makers are **households** (or consumers) and **firms** (or businesses); both are shown in square boxes. Households and firms are linked together through two markets: product markets and resource markets, shown in diamonds.

Households are owners of the four factors of production: land, labour, capital and entrepreneurship. Firms buy the factors of production in resource markets and use them to produce goods and services. They then sell the goods and services to consumers in product markets. We therefore see a flow in the clockwise direction of factors of production from households to firms, and of goods and services from firms to households.

In the counterclockwise direction, there is a flow of *money* used as payment in sales and purchases. When households sell their factors of production to firms,

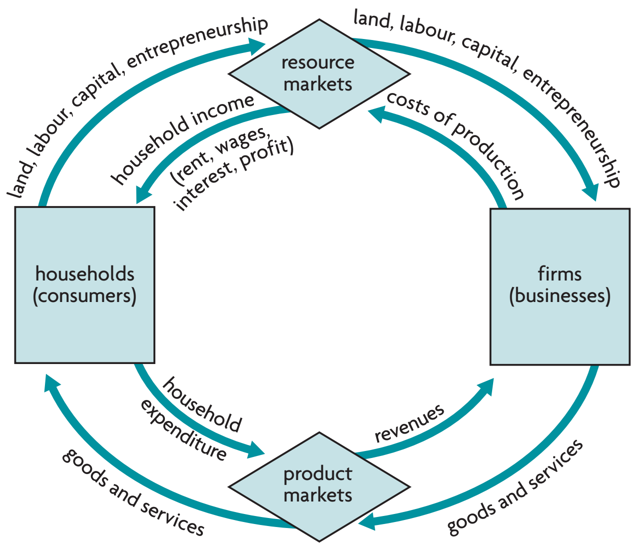


Figure 1.4: Circular flow of income model in a closed economy with no government

they receive payments taking the form of *rent* (for land), *wages* (for labour), *interest* (for capital) and *profit* (for entrepreneurship). These payments are the *income of households*. The payments that households make to buy goods and services are *household expenditures* (or consumer spending). The payments that firms make to buy factors of production represent their *costs of production*, and the payments they receive by selling goods and services are their *revenues*. All payment flows, known as *money flows*, are shown in Figure 1.4

This model demonstrates an important principle: the *income flow* involving the money that goes from firms to households is equal to the *expenditure flow* from households to firms, which is the money that households spend to buy things from firms. In other words, the household incomes coming from the sale of all the factors of production equals the expenditures by households on goods and services. This is the *circular flow of income*.

In addition, these two flows must be equal to the value of goods and services, or the value of total output produced by the firms, known as the *value of output flow*. The reasoning of this is as follows: if each good and service is multiplied by its respective price, we obtain the value of each good and service, and adding them all up we arrive at the value of total output. This value is the same as consumer expenditure, since spending by consumers is equal to each item they buy multiplied by its price. Therefore:

The **circular flow of income** shows that in any given time period (say a year), the value of output produced in an economy is equal to the total income generated in producing that output, which is equal to the expenditures made to purchase that output.

Adding leakages and injections

The real-world economy is more complicated than this simple model suggests. We arrive at a closer picture of the real world by adding **injections** and **leakages** (also known as *withdrawals*) to the money flow of Figure 1.4. To understand what these are, consider a pipe with water flowing through it, as in Figure 1.5. As water flows through the pipe, some leaks out (the leakages), while new supplies of water are injected in (the injections). It is the same with the flows of money in the circular flow model.

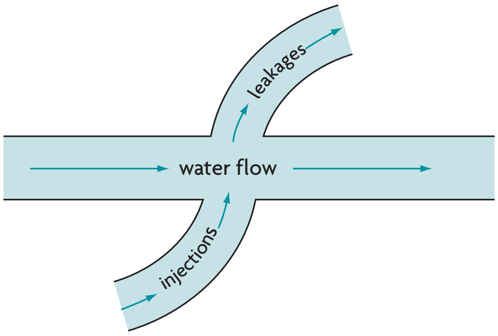


Figure 1.5: Leakages and injections

Leakages and injections are paired together so that what leaks out of the flow can come back in as an injection. The most important pairs are the following:

leakages	injections
saving	investment
taxes	government spending
imports	exports

Saving and investment

Saving is the part of consumer income that is not spent but rather is saved. Investment is spending by firms for the production of capital goods, which is one of the four factors of production. This is why capital goods are also known as investment goods. How are saving and investment linked together as leakages and injections?

When households save part of their income, this represents a leakage from the circular flow of income because it is income that is not spent to buy goods and services. Households place their savings in financial markets (bank accounts, purchases of stocks and bonds, etc.). Firms obtain funds from financial markets (through borrowing, issuing stocks and bonds, etc.) to finance investment, or the production of capital goods. These funds therefore flow back into the expenditure flow as injections. This process is shown in Figure 1.6, which, in addition to the money flows of Figure 1.4, shows the three leakage/injection pairs. (For simplicity, Figure 1.6 contains only money flows.) Leakages appear in the left-hand side of the figure, and injections on the right. We can see that saving leaks out of the flow of consumer expenditures (saving is money that is not spent), and after passing through financial markets is injected back into the expenditure flow as investment.

Taxes and government spending

Taxes and government spending are connected to each other through the government. Households pay taxes to the government; this is a leakage because it is income that is not spent to buy goods and services. The government uses the tax funds to finance government expenditures (on education, health, defence, etc.) and this spending is an injection back into the expenditure flow.

Imports and exports

Imports are goods and services produced in other countries and purchased by domestic buyers. **Exports** are goods and services produced domestically and purchased by foreigners. When an economy has international

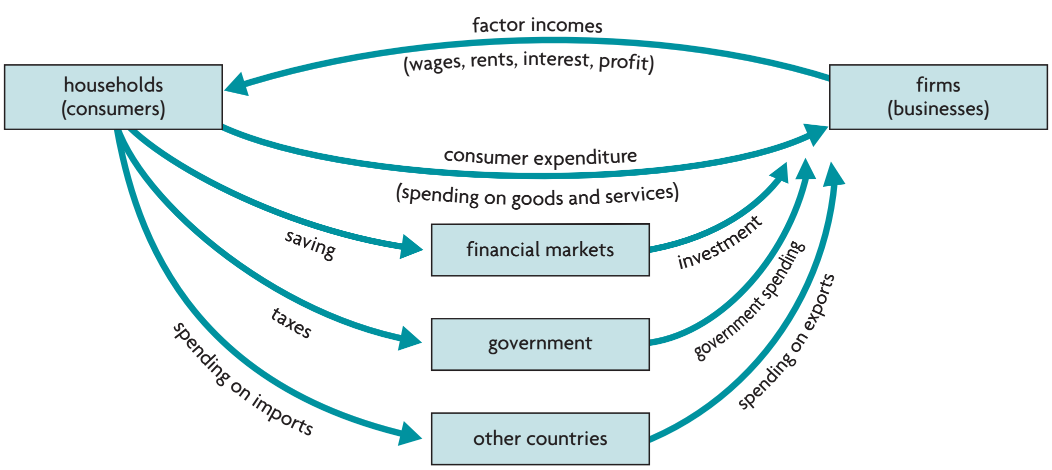


Figure 1.6: Circular flow of income model with leakages and injections

trade with imports and exports, it is known as an *open economy*. Imports and exports are linked together through 'other countries'. Imports are a leakage because they represent household spending that leaks out as payments to the other countries that produced the goods and services. Exports are an injection because they are spending by foreigners who buy goods and services produced by the domestic firms.

The size of the circular flow in relation to the size of leakages and injections

In the real world, leakages and injections are unlikely to be equal, and this has important consequences for the size of the circular flow. If a leakage is greater than an injection, then the size of the circular flow becomes smaller. Suppose saving (a leakage) is larger than investment (an injection). This means that part of the household income that leaks as saving into financial markets does not come back into the flow as investment. The result is that fewer goods and services are purchased, firms cut back on their output, they buy fewer factors of production, unemployment increases (since firms buy a smaller quantity of labour) and household income is reduced.

If a leakage is smaller than an injection, the size of the circular flow becomes larger. Suppose spending on exports is greater than spending on imports; then the expenditure flow increases since the injection is larger than the leakage. Foreigners demand more goods and services, firms begin to produce more by purchasing more factors of production, unemployment falls (as firms buy a larger quantity of labour), and household income increases. To summarise, leakages from the circular flow of income (saving, taxes and imports) are matched by injections into the circular flow of income (investment, government spending and exports), though *these need not be equal to each other*.

In the *circular flow of income model* if injections are larger than leakages the size of the flow increases; if leakages are larger than injections the size of the flow shrinks.

TEST YOUR UNDERSTANDING 1.10

1. a. Identify the two markets shown in the circular flow of income model. Using examples, outline what is exchanged (bought and sold) in each of these.
   b. Identify the three flows shown in the circular flow model.
2. The circular flow of income model shows that households and firms are both buyers and sellers simultaneously. Explain how this is possible.
3. Use the simple circular flow of income model to:
   a. show the circular flow of income
   b. show the equivalence between factor income flow, household expenditure flow and the value of output flow.
4. a. Define leakages and injections in the circular flow of income.
   b. Use the circular flow of income model to illustrate how the three pairs of leakages and injections are linked together.
5. Distinguish between a closed economy and an open economy.
6. Describe what happens to the size of the income flow when:
   a. leakages are larger than injections
   b. injections are larger than leakages.

The role of microeconomics and macroeconomics in the circular flow model

The circular flow of income model is a simple model that *describes* basic economic relationships, bringing together and showing how microeconomics and macroeconomics relate to each other.

(Microeconomics and macroeconomics were introduced at the beginning of this chapter. We can see in this model that the basic decision-making units, *consumers* and *firms*, are interdependent; they are linked together through their buying and selling activities which occur in markets. We will study the behaviour of consumers and firms in *microeconomics*. These buying and selling activities, when added all up, lead to flows of income, output and spending, which are also interdependent. We will study these under *macroeconomics*.

1.4 The method of economics

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between positive and normative economics (AO2)
- explain the use of logic, hypotheses, models and theories in positive economics (AO2)
- explain the role of the *ceteris paribus* assumption (AO2)
- explain the roles of empirical evidence and refutation in positive economics (AO2)
- explain the role of value judgements in policy-making in normative economics (AO2)
- distinguish between equity and equality (AO2)

Positive versus normative economics

Explaining positive and normative statements

Economists think about the economic world in two different ways:

i One way of thinking tries to describe, explain and predict economic events; this is called *positive economics*. It is based on *positive statements,* which are about something that is, was or will be. Positive statements are used to describe, explain or predict economic events by use of hypotheses, theories and models. Positive statements:

- may *describe* something; for example, 'the unemployment rate is 5%' and 'industrial output grew by 3%' are two statements describing the level of unemployment and growth of industrial output
- may be statements in a hypothesis that tries to *explain* something; for example, the statement 'a higher price of apples results in fewer apples purchased' is a statement that provides an explanation of why fewer apples are purchased
- may be statements that *predict* a future event; for example, 'unemployment will increase next year' predicts what will happen to unemployment next year.

ii The other way of thinking deals with how things in the economy should or ought to be; this is called *normative economics*. It is based on beliefs or value judgements about what *should happen*, about *what is good or bad*, about *what is right or wrong*. It is used in making economic policies. Examples include the following:

- 'The unemployment rate should be lower' is a statement based on a belief or value that high unemployment is not a good thing.
- 'Health care should be available free of charge' is a statement about a value that everyone in a society should have access to free health care.
- 'The government should spend more money on building schools' is a policy recommendation about what the government should do.

See Table 1.3 for more examples of positive and normative economics. Note that statements in positive economics may be factually correct or they may be false. For example, we may say that the unemployment rate is 5%; if the unemployment rate is 5%, this statement is correct; but if the unemployment rate is actually 7%, the statement is false.

Statements in normative economics, by contrast, cannot be true or false. They can only be assessed relative to *beliefs* and *value judgements*. Consider the normative statement 'the unemployment rate should be lower'. We cannot say whether this statement is true or false, though we may agree or disagree with it, depending on our beliefs about unemployment. If we believe that the present unemployment rate is too high, then we will agree; but if we believe that the present unemployment rate is not too high, then we will disagree.

Positive economics	Normative economics
Incomes have fallen by 7%.	Incomes should be increased.
Free university education will increase government spending by 3%.	There should be free university education.
Income inequality is increasing.	Income inequality has increased too much.
Women are often paid less than men for the same work.	Women and men should receive equal pay for the same work.
Higher taxes will result in lower disposable incomes.	Taxes are too low and should be increased.

Table 1.3: More examples of positive and normative economics

TEST YOUR UNDERSTANDING 1.11

1 Which of the following are positive statements and which are normative?

- a It is raining today.
- b It is too humid today.
- c Economics is a study of choices.
- d Economics should be concerned with how to reduce poverty.
- e If household saving increases, there will be a fall in household spending.
- f Households save too little of their income.

2 Explain the importance of making a distinction between positive and normative statements in economics.

The role of positive economics

In economics, as in other social (and natural) sciences, our efforts to gain knowledge about the world involve the formulation of hypotheses, theories, laws and models. All of these are based on the use of logic. All of these lie within the realm of positive economics.

The use of logic

As we have seen above, positive economics involves thinking about the economic world in order to try to describe, explain or make predictions about economic events. Positive economic thinking is used to describe and explain in a systematic way why economics events happen the way they do, and attempts to predict economic events that are likely to occur in the future.

This type of thinking is based on the use of **logic**, a Greek word (λογική or logiki) which means *reason*. It is a *method of reasoning*, which involves making a series of statements each of which is true if the preceding statements are true. For example:

1 When it rains there are clouds in the sky.
2 It is now raining.
3 Therefore there are clouds in the sky.

The truth of the third statement is based on the truth of the previous statements. Therefore, we can say that the third statement is logical, as it is based on logic.

Economists use logical thinking to acquire knowledge of the economic world. Since economics is a social *science*, economists acquire knowledge by use of the scientific method (you may already be familiar with this from your study of the natural sciences like physics, biology and chemistry). The scientific method is based on the use of logic.

The use of hypotheses

A **hypothesis** is an educated guess, usually indicating a cause-and-effect relationship about an event. Hypotheses are often stated as: if …, then …

In order to formulate a hypothesis, economists make observations of the world around them and identify a question they would like to answer. Let's consider an example from economics. We observe that people living in the city of Olemoo buy different amounts of oranges per week at different times in the year. We want to answer the question: why are more oranges bought in some weeks and fewer in others?

We then identify variables that we think are important to answer the question. A variable is any measure that can take on different values, such as temperature, or weight or distance. In this example, the variables we choose to study are the quantity of oranges that residents of Olemoo buy each week, and the price of oranges.

Our next step is to make a *hypothesis* about how the variables are related to each other. We have seen that a hypothesis is an educated guess about an *if . . . then . . .* relationship. Our hypothesis is the following: ***if** the price of oranges increases, **then** the quantity of oranges Olemooans want to buy each week will fall*. Notice that this hypothesis indicates a cause-and-effect relationship, where price is the 'cause' and the quantity of oranges bought is the 'effect'. The hypothesis also involves a prediction, because it claims that changes in the price of oranges will lead to a particular change in the quantity of oranges Olemooans buy.

The ceteris paribus assumption

If our hypothesis is to make sense, we need to also make assumptions. An *assumption* is a statement that is supposed to be true for the purposes of building the hypothesis. In our example, an important assumption is this: the price of oranges is the only variable that influences the quantity of oranges that Olemooans want to buy, while all other variables that could have influenced their buying choices do not play a role. This is called the *ceteris paribus* assumption. *Ceteris paribus* is a Latin expression that means 'other things equal'. Another way of saying this is that all other things are assumed to be constant or unchanging.

Why is this assumption important? Our hypothesis stated that the quantity of oranges that will be bought is determined by their price. Surely, however, price cannot be the only variable that influences how many oranges Olemooans want to buy. What if the population of Olemoo increases? What if the incomes of Olemooans increase? And what if an advertising campaign proclaiming the health benefits of eating oranges influences the tastes of Olemooans? As a result of any or all these factors, Olemooans will want to buy more oranges.

This complicates our analysis, because if all these variables change at the same time, we have no way of knowing what effect each one of them individually has on the quantity people want to buy. We want to be able to isolate the effects of each one of these variables; to test our hypothesis we specifically wanted to study the effects of the price of oranges alone. This means we have to make the assumption that all other things that could affect the relationship we are studying must be constant, or unchanging.

More formally, we would say that we are examining the effect of orange prices on the quantity of oranges people want to buy, *ceteris paribus*. This means simply that we are studying the relationship between prices and quantity *on the assumption* that nothing else happens that can influence this relationship. By eliminating all other possible interferences, we isolate the impact of price on quantity, so we can study it alone.

In the real world all variables are likely to be changing at the same time. The *ceteris paribus* assumption does not say anything about what happens in the real world. It is simply a tool used by economists to construct hypotheses, models and theories, thus allowing us to isolate and study the effects of one variable at a time. We will be making extensive use of the *ceteris paribus* assumption in our study of economics.

The use of empirical evidence

Now we are in a position to test our hypothesis to see if its predictions fit with what actually happens in the real world. To do this, we compare the predictions of the hypothesis with real-world events, based on **empirical evidence**. Empirical evidence refers to real-world information, observations and data that we acquire through our senses and experience (*empirical* comes from the Greek word εμπειρία or empeiría meaning experience).

Here, the methods of economics differ from those of the natural sciences. Whereas in the natural sciences it is often (though not always) possible to perform experiments to test hypotheses, in economics the possibilities for experiments are very limited. Economists therefore rely on a branch of statistics called econometrics to test hypotheses. This involves collecting data on the variables in the hypothesis, and examining whether the data fit the relationships stated in the hypothesis. In our example, we must collect data on the quantity of oranges bought by Olemoo's residents during different weeks throughout the year, and compare these quantities with different orange prices at different times in the year. (Econometrics is usually studied at university level, and is not part of IB requirements.)

We are now in a position to compare the predictions of our hypothesis with real-world outcomes. If the data did not fit the predictions of the hypothesis, the hypothesis would be rejected, and the search for a new hypothesis would begin. In our example, this would happen if we discovered that as the price of oranges increases, the quantity of oranges Olemooans want to buy each week also increases. Clearly, this would go against our hypothesis, and we would have to reject the hypothesis as invalid. If, on the other hand, the data fit the predictions, the hypothesis would be accepted. In our example, this would occur if our data show that as the price of oranges increases, Olemoo's residents buy fewer oranges. We can therefore conclude that according to the evidence, our hypothesis is a valid one.

Theories in relation to hypotheses

We have seen that a hypothesis is an educated guess about a cause-and-effect relationship in a single event. A **theory** is a general explanation of a set of interrelated events, usually (though not always) based on several hypotheses that have been tested successfully (in other words, they have not been refuted, or disproven, based on evidence; see the discussion below on *refutation*). A theory is a generalisation about the real world that attempts to organise complex and interrelated events and present them in a systematic and coherent way to explain *why* these events happen. Based on their ability to systematically explain events, theories attempt to make predictions.

Referring to the example of oranges, the relationship between the price of oranges and the quantity of oranges residents of Olemoo buy at each price was a hypothesis. This kind of hypothesis has been successfully tested a great many times for many different goods, and the data support the presence in the real world of such a relationship. However, this relationship is not a theory, because it only shows how two variables relate to each other, and does not explain anything about *why* buyers behave the way they do when they make decisions to buy something.

To explain this relationship in a general way, economists have developed 'utility theory' and 'indifference curve analysis' based on a more complicated analysis involving more variables, assumptions and interrelationships. These theories try to answer the question as to *why* people behave in ways that make the observed relationship between price and quantity a valid one. (Utility theory will be examined at HL in Chapter 2).

The use of laws

A *law*, in contrast to a theory, is a statement that describes an event in a concise way, and is supposed to have universal validity; in other words, to be valid at all times and in all places. Laws are based on theories and are known to be valid in the sense that they have been successfully tested very many times. They are often used in practical applications and in the development of further theories because of their great predictive powers. However, laws are much simpler than theories, and do not try to explain events the way theories do.

For example, the simple relationship between the quantity of a good that people want to buy and its price, while not a theory, has the status of one of the most important *laws* of economics: it is the *law of demand*. This law is a statement describing an event in a simple way. It has great predictive powers and is used as a building block for very many complex theories. We will study the law of demand in detail in Chapter 2 and we will use it repeatedly throughout this book in numerous applications, and as a building block for many theories.

The use of models

In your study of economics, you will encounter many theories and some laws. Your study of both theories and laws will make great use of economic models. Models are sometimes used to illustrate theories (or laws) and sometimes to describe the connections between variables.

Whereas sciences like biology, chemistry and physics offer the possibility to construct three-dimensional models (as with molecules and atoms), this cannot be done in the social sciences, because these are concerned with human society and social relationships. In economics, models are often illustrated by use of diagrams showing the relationships between important variables. In more advanced economics, models are illustrated by use of mathematical equations. (Note that both diagrams and mathematical equations are used to represent models in natural sciences, such as physics, as well.)

Models are often closely related to theories, as well as to laws. A theory tries to explain *why* certain events happen and to make predictions. A law is a concise statement of an event that is supposed to have universal validity. Models are often built on the basis of well-established theories or laws, in which case they may illustrate, through diagrams or mathematical equations, the important features of the theory or law. When this happens, economists use the terms 'model' and 'theory' interchangeably because in effect they refer to one and the same thing. For example, in Chapter 9, different models of the macroeconomy will be used to illustrate alternative theories of income and output determination.

However, models are not always representations of theories. In some cases, economists use models to isolate important aspects of the real world and show connections between variables but without any explanations as to *why* the variables are connected in some particular way. In such cases, models are purely descriptive; in other words, they describe a situation without explaining anything about it. For example, the production possibilities curve model presented above in Section 1.3 is a simple model that is very important because of its ability to *describe* scarcity, choice and opportunity cost. Similarly, the circular flow of income model, also presented in Section 1.3, *describes* how decision-makers are related to each other in the economy and introduces the concepts of output, income and spending. In the case of both models, there is no theory or explanation involved.

Yet descriptive models that are not based on a theory are in no way less important than models that illustrate a theory. Both kinds of models are very effective as tools used by economists to highlight and understand important relationships and phenomena in the economic world. In our study of economics, we will encounter a variety of economic models and will make extensive use of diagrams.

The importance of refutation

The concept of **refutation** is very important in economics (as in any social science or science). To *refute* something means to contradict it, disprove it or show it to be false. Refutation in the sciences and social sciences is the idea that it must be possible to *refute or disprove* a hypothesis or a theory. It must be possible to subject it to empirical testing, where the data or empirical observations can disprove it if it is false or invalid. In other words, if a hypothesis or theory cannot be refuted or disproven by empirical testing, then it is not scientific. Refutation is also known as *falsifiability*, because if something is refuted, it is falsified, in other words it is shown to be false.

You may note that our hypothesis about Olemooans is refutable or falsifiable. The hypothesis is 'if the price of oranges increases, then the quantity of oranges Olemooans want to buy each week will fall'. We could collect data on quantities and prices of oranges, and if the data do not fit the hypothesis, the hypothesis would be *refuted*, in other words it would be falsified or disproven.

TEST YOUR UNDERSTANDING 1.12

1. The relationships between hypotheses, theories, laws and models described here apply generally to all the social sciences and sciences based on the scientific method. Yet they may differ between disciplines in the ways they are used and interpreted. As you study economics, you may want to think about the following: How are theories and laws used in economics as compared with other disciplines? Do they play the same role? Are they derived in the same ways? Do they have the same meaning?
2. Explain the scientific method. Outline the steps it involves.
3. Distinguish between hypotheses, theories, laws and models.
4. a. Explain why it is important to compare the predictions of a hypothesis with real-world outcomes.
   b. Explain the role of empirical evidence and refutation in the scientific method.
5. Describe how models help economists in their work as social scientists.
6. Consider the statement, 'If you increase your consumption of calories, you will put on weight.'
   a. Explain why this statement is not necessarily true.
   b. Rephrase the statement to make it more accurate. (*Hint:* What might happen to your weight if at the same time that you consume more calories you also began an intensive exercise programme?)

THEORY OF KNOWLEDGE 1.1

Refutation, science and truth

We have seen how hypotheses are tested using the scientific method. If the data support the predictions of a hypothesis, the hypothesis is accepted. However, this does not make the hypothesis necessarily 'true'. The only knowledge we have gained is that *according to the data used, the hypothesis is not false*. There is always a possibility that as testing methods are improved and as new and possibly more accurate data are used, a hypothesis that earlier had been accepted now is rejected as false. Therefore, no matter how many times a hypothesis is tested, we can never be sure that it is 'true'.

But by the same logic, we can never be sure that a hypothesis that is rejected is necessarily false. It is possible that our hypothesis testing, maybe because of poor data or poor testing methods, incorrectly rejected a hypothesis. Testing of the same hypothesis with different methods or data could show that the hypothesis had been wrongly rejected.

If our results from hypothesis testing are subject to so many uncertainties, how can economic knowledge about the world develop and progress? Economists and other social and natural scientists work with *hypotheses that have been tested and not refuted* (not falsified or disproven). While the possibility exists that the hypotheses may be false, they use these hypotheses *on the assumption* that they are not false. As more and more testing is done, and as unfalsified hypotheses accumulate, it becomes more and more likely that they are not false, *though we can never be sure*.

This way, it is possible to accumulate knowledge about the world, however, this is done on the understanding that this knowledge is tentative and provisional; in other words, it can never be proven to be true.

Thinking points

- Is it possible to ever arrive at the truth of a statement about the real world based on empirical testing?
- Even assuming that testing methods could be perfected and data vastly improved, can there ever be complete certainty about our knowledge of the social (and natural) worlds?

The role of normative economics

Value judgements in policy-making

It was noted earlier that normative economics is based on beliefs and *value judgements* about what should happen, about what is good or bad, or about what is right or wrong. Value judgements are opinions; they are subjective judgements rather than factual statements.

Note that statements based on normative economics are not refutable or falsifiable. They cannot be shown to be false. One can only agree or disagree with them depending on one's own beliefs and value judgements.

Value judgements in normative economics are important for economic policy-making. They identify the important economic problems that should be addressed and recommend policies to solve them. Economic policies are government actions that try to solve economic problems. Examples of economic policies are government actions to lower unemployment, lower inflation, protect the environment, improve the quality of education, reduce levels of poverty, provide free health care services and many more. When a government makes a policy to lower unemployment (the number of people who are looking for work but can't find a job), this is based on the value judgement that high unemployment is not a good thing. If a government pursues a policy to make health care available to everyone free of charge, this is based on a value judgement that people should not have to pay for receiving health care services.

Positive economics and normative economics, while distinct, often work together. To be successful, an economic policy aimed at lowering unemployment (the normative dimension) must be based on a body of economic knowledge about what causes unemployment (the positive dimension). The positive dimension can provide guidance to policy-makers on how to achieve their economic goals.

Positive economics is the study of economics based on the scientific method, used to arrive at knowledge about the economic world. It includes descriptions, models, hypotheses, theories and laws. **Normative economics** forms the basis of judgements about what economic goals and economic policies ought to be. It is based on value judgements, because it identifies important economic problems that should be addressed and prescribes what should be done to solve them.

Equity and equality

An important concept from normative economics that you will encounter in this book is *equity,* which refers to the idea of being fair or just. The idea of equity, or fairness is a normative concept because fairness depends on people's beliefs or value judgements, which differ from person to person.

Equality, on the other hand, is the state of being equal with respect to something. For example, equality with respect to income would mean that each member of a society receives exactly the same income. Equality is a positive concept, since two or more things are either equal or not equal based on a measure.

The idea of equality in income distribution may or may not be equitable, depending on how equity is interpreted. If it is believed that income distribution is equitable or fair if income is distributed equally, then equity in income distribution means income equality. However, if it is believed that it is equitable or fair for people's income to be in proportion to their work effort (a different equity principle), this would give rise to income inequality, since not everyone's work effort is the same.

In spite of different possible interpretations of the meaning of equity, in most countries around the world, the pursuit of equity is usually interpreted to refer to government policies that try to reduce inequalities in income, wealth and opportunity. This is because of the widely shared belief or value judgement that the free market without government intervention results in highly unequal income and wealth distributions that are considered to be unfair. For this reason, we often find writers referring to a 'more equitable' or 'more equal' distribution of income or wealth to mean the same thing. Both expressions are correct, provided it is understood that in these cases, *equity in income distribution is interpreted as greater equality (or less inequality)*.

Equity means fairness while **equality** means being the same. While the two concepts have different meanings, in economics **inequity** is most often interpreted to mean **inequality**, while equity is interpreted to mean equality.

Note that the pursuit of equity in the sense of equality does not refer to efforts to achieve complete income or wealth equality, but rather the reduction of inequalities that are considered to be unfair.

You should also note that the ideas of equity and equality in distribution involve answers to the *for whom to produce* basic question of economics (see Section 1.2). The other two questions, *what/how much to produce* and *how to produce* are answered by use of positive economics.

TEST YOUR UNDERSTANDING 1.13

1 Suppose the government of Country X implements a series of policies to reduce income inequalities. Identify possible value judgements that may have led to this policy.

2 Distinguish between equity and equality.

THEORY OF KNOWLEDGE 1.2

Why do economists often disagree?

As you will discover in the course of reading this book, there are many areas of economic theory as well as policy on which there are major disagreements and debates among economists. Why do economists disagree so much? It would seem that use of the scientific method in economics, by forcing hypotheses to undergo tests, and allowing the real-world evidence to sift through valid and invalid hypotheses, would eliminate much disagreement. Why do economists continue to disagree in spite of their use of the

CONTINUED

scientific method? To try to answer this question, we should consider the point mentioned earlier on the difficulties of testing hypotheses due to the inability of economists to perform controlled experiments.

The scientific method, as we have seen, involves relating evidence to educated guesses about cause-and-effect relationships between variables to see if they match. Economists face some difficulties in this effort. First, the inability to perform controlled experiments means that economists collect data about real-world events that are the result of many variables changing at the same time. To test hypotheses, economists devise complicated econometric models that try to isolate the interfering effects of numerous variables, and try to link causes with effects. Sometimes, economists have to deal with incomplete or unreliable real-world data. In some cases, they may even be faced with variables that are not measurable and have no data, in which case they must use substitute variables (called 'proxy' variables) or substitute relationships between variables. As a result of these difficulties, it is not unusual for two or more economists to be testing the same hypothesis and to come up with conflicting results.

For all these reasons, while the testing methods of economists do produce some useful results, these are sometimes not as accurate and as reliable as the results of experiments in other disciplines performed under controlled conditions. This means it may be more difficult for hypothesis testing in economics to refute (reject) invalid hypotheses. If the evidence does not reject a hypothesis, economists hold on to it and may continue to use it in their work (possibly until further testing in the future). However, this does not mean that the hypothesis is a valid one. It may be invalid, but the evidence just has not been discriminating enough to reject it. This has important implications for economics. It means that there may be several conflicting hypotheses that economists are holding on to and working with, *not all of which are valid hypotheses*, and some of which may be false.

Moreover, economists may use these hypotheses to build theories. A theory was described earlier as being based on several hypotheses that have not been rejected, based on evidence. This means it is possible to have theories built on invalid hypotheses, which simply have not (yet) been shown to be invalid. But if the hypotheses on which theories are built are invalid, then surely the theories themselves are also invalid. This explains one possible reason why we sometimes see several conflicting theories being used at the same time. Maybe only one of them (or even none of them) is valid. Whatever the case, as economists usually prefer to support one theory over another, this may be an important reason why they sometimes disagree.

As we will see in Theory of knowledge 9.1 (Chapter 9), the inability of empirical evidence to effectively discriminate between competing theories allows much room for value judgements and ideologies to enter into one's individual preference for one economic theory over another.

Thinking points

As you read this book and learn more about economics, you may want to keep the following questions in mind:

- Can you think of other possible reasons why economists often disagree?
- What other social sciences/sciences cannot test hypotheses by performing controlled experiments?
- Do you think economists disagree more or less than (or the same as) other social and natural scientists?
- Do you think the difficulties of economics are due to its being a 'young' social science that will slowly 'mature' and resolve these difficulties as econometric methods and the quality of data improve? Or are they due to problems that are inherent in the nature of the subject and therefore cannot be easily resolved?
- Do you think these difficulties seriously affect the progress and development of new economic knowledge? Or can economics continue to progress in spite of these difficulties?

1.5 A brief history of economic thought: the origins of economic ideas

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the major schools of thought from the 18th century to the present: (AO2)

  18th century:

  - Adam Smith and laissez faire

  19th century:

  - utility theory in classical microeconomics
  - the concept of the margin
  - Say's Law in classical macroeconomics
  - the Marxist critique of classical economics

  20th century:

  - the Keynesian revolution
  - the emergence of macroeconomic policy
  - the monetarist/new classical counter-revolution

  21st century:

  - behavioural economics and the dialogue with psychology
  - growing awareness of the interdependence between the economy, society and the environment, and the need to move toward a circular economy

The history of economic thought is a fascinating account of how economic ideas have evolved over the years. Many of the economic ideas and theories you will learn about in your study of economics can be traced back to the contributions of famous economists who lived and worked decades, and even centuries, ago. Although economic thinking has developed and progressed over time, leading to a deeper understanding of economic events, many of these contributions still lie at the heart of economics that we study today.

It is strongly recommended that you read this section again after completing the rest of this book. It is only then that you will be able to appreciate the richness of the ideas of these famous economists, because only then will you be able to recognise in their work your own understanding of economics.

Until about the 18th century, there were no distinct disciplines as we know them today. Investigation of events and phenomena in the social world were part of what was known as *moral philosophy*, while investigation of events and phenomena in the natural world were part of what was known as *natural philosophy*.

In the area of what we now know as *economics*, scholars and philosophers since the time of the ancient Greeks, two and a half thousand years ago, have concerned themselves with ideas that try to explain economic events, but no one had attempted to provide a theory explaining the economy in a systematic way. The first scholar to do this was Adam Smith.

The 18th century: Adam Smith and laissez faire

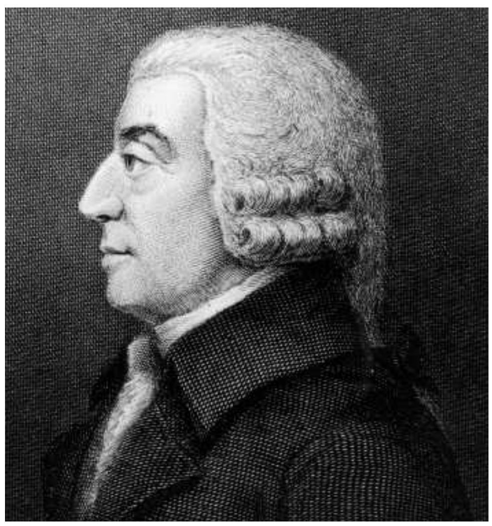


Figure 1.7: Adam Smith, Scottish philosopher and political economist, often considered to be the 'Father of Economics', author of *An Inquiry into the Nature and Causes of the Wealth of Nations* (the *Wealth of Nations* for short)

Adam Smith (1723–1790) was a Scottish philosopher who is often referred to as the *Father of Economics*. Adam Smith studied moral philosophy at the University of Glasgow and at Oxford University, eventually ending up back at the University of Glasgow, where he taught moral philosophy. He is best known for the ideas expressed in his book *The Wealth of Nations*,[5] published in 1776.

The Wealth of Nations is considered to be a masterpiece that has influenced economics for generations up to the present day. Although it borrows heavily from the ideas of various scholars of the time, it represents the first attempt ever to set forth a comprehensive theory of how an economy can hold itself together, functioning in a harmonious way, and how such an economy can grow over time.

Adam Smith lived at a time when the traditional rights of rulers to impose authoritarian controls and restrictions on their subjects in Europe was coming more and more under attack. Smith believed that strong, repressive governments were not essential to the workings of an economy. He therefore set out to show how an economic system without government could not merely function, but could moreover thrive and prosper to the great benefit of its citizens.

In this task, Smith borrowed heavily from the ideas of the natural laws set forth by Isaac Newton (1642–1727), the great mathematician, physicist, astronomer and theologian. Just as Newton discovered natural laws that govern the physical world with harmonious regularity, so Smith believed he was uncovering the natural laws of the social world that govern economic relations with that same regularity.

Smith begins with the idea of a market and the behaviour of individuals in a market that will produce beneficial results for the whole of society. He shows with painstaking detail and numerous examples in *The Wealth of Nations* that if individuals behave in a self-interested way, so that each person tries to do the best for herself or himself, there will result a greater good for society. In a famous quote from *The Wealth of Nations*, Smith writes:

> *'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.'*

Such behaviour gives rise to *competition*, which regulates the behaviour of people acting in their self-interest. Competition is very important because it keeps people's self-interest in check. Suppose, for example, that a producer of shoes raises price to a high level expecting more personal gain. Other producers of shoes who also want to sell their products will sell them at a lower price, so the high-price seller will be forced to lower the price. This process continues until the price of shoes falls to the lowest possible level.

Similarly, competition will lead to the production of those goods and services that are mostly wanted, because self-interest will lead producers to make those goods that they will be able to sell. In the same way, occupations that are difficult or dangerous will command higher wages because workers will only be willing to take on such jobs if the wages are high enough to compensate them for the hard work and risks.

These ideas gave rise to the famous expression the *invisible hand of markets*, which refers to the presence of thousands or millions of individual decisions made by individual decision-makers that are self-regulated through interactions in markets, without the presence of a government deciding what/how much to produce. The invisible hand results in a more efficient use of resources.

This is also known as a **laissez faire** economy (from French, meaning *let it do*) which refers to a free market without government intervention. It is not necessary to have a government telling people what to do or how to do it, because the market working on its own can do a much better job of this.

However, Adam Smith did foresee a role for government as well. Governments in his view have three important functions: to take care of national defence, to oversee security and a system of justice, and to provide public infrastructure (roads, bridges, canals and so on), which are essential to the proper functioning of an economy and therefore should be financed from taxation.

The idea of competition, being central to Smith's ideas, was carried further. Smith was very concerned about the possible growth of firms to become large corporations or monopolies, which would restrict competition. He was keenly aware of the power of large firms to raise the prices to high levels and keep them there due to the absence of competition. As he wrote:

> *'The price of monopoly is upon every occasion the highest which can be got. The natural price, or the price of free competition, on the contrary, is the lowest which can be taken.'*

[5] The full title of the book was *An Inquiry into the Nature and Causes of the Wealth of Nations.*

Smith believed that competition would ensure that monopolies and large corporations will not be able to arise.

The Wealth of Nations also presents a theory of economic growth. Economic growth is seen to depend on the division of labour (the separation of processes into many different tasks, each one performed by different workers) which permits *specialisation* of labour to take place. Specialisation of labour involves the pursuit by workers of a particular task involving skills appropriate to the task in question. If workers specialise in particular activities that they can perform best, then more output will be produced and this will lead to economic growth.

Smith extended the idea of specialisation beyond what takes place within countries to international trade between countries. He is responsible for the idea of *absolute advantage* of a country in international trade. According to this idea if each country produces the goods it can produce at the lowest cost and trades them for goods produced at a lower cost in other countries, then all the countries involved will be better off. Therefore, Adam Smith was an advocate of free trade (trade without restrictions). The theory of absolute advantage will be discussed in Chapter 14 (at HL only).

Whereas Smith wrote that self-interested behaviour leads to the beneficial invisible hand of the market, he did not write that people's actions are determined only by self-interest. In an earlier book, *The Theory of Moral Sentiments* (1759), Smith argued that people have the ability to put themselves in the position of another person, in other words to empathise. This makes them behave in ways that create happiness for other people, because doing so gives them pleasure. This is important because in later years Smith's ideas were misinterpreted to mean that people behave only selfishly and with greed.

The concepts of demand, supply and markets that you will read about in Chapter 2 are extensions and elaborations of Adam Smith's thinking.

> Adam Smith is best known for the idea that the self-interested behaviour of decision-makers without government intervention results in competitive markets that give rise to a more efficient use of resources and greater output, thus benefitting society. This is known as the *invisible hand of the market.*

The 19th century

The economic ideas that developed during the 19th century are known as **classical economics**. Economic thinking was modified and refined considerably since the time of Adam Smith. The main economists of the time were concerned with issues like the process of economic growth and the distribution of income. A number of scholars made important contributions, such as David Ricardo (who we will encounter in Chapter 14), and Jeremy Bentham and John Stuart Mill who we will briefly discuss below.

Utility theory in classical microeconomics

Jeremy Bentham (1748–1832) and John Stuart Mill (1806–1873) were British philosophers who also made contributions to economics. Jeremy Bentham was the founder of *utilitarianism*, a philosophy of ethics (a philosophy about what is right and wrong) which taught that an action is right if it promotes the most happiness for the largest number of people. Bentham's theory of ethics is based on the idea that:

> *'it is the greatest happiness of the greatest number that is the measure of right and wrong.'*

Therefore actions are right or wrong according to the consequences they have on the happiness of other people.

John Stuart Mill furthered Bentham's ideas by blending them with human rights, including the rights of minorities and women. He was a fervent believer in human freedom and was opposed to forcing people to do things against their will. Mill's theory of ethics can be summarised in his statement that:

> *'actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.'*

The idea of the 'greatest happiness' is based on the concept of *utility*, which Bentham defined to be:

> *'that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness . . . or . . . to prevent the happening of mischief, pain, evil, or unhappiness.'*

Similarly, Mill defines happiness as pleasure and the absence of pain.

Utility, the central concept in the philosophy of ethics of utilitarianism, evolved to become a central concept of economics that underlies economic theories up to the present day.

Classical economists developed the philosophy of ethics known as *utilitarianism*, according to which an action is right if it promotes the most happiness for the largest number of people.

The concept of the margin

Classical economists were concerned about the concept of *value*, specifically what gives things their value and what determines their price. One theory that was popular at the time was the labour theory of value, according to which the price of a good was determined by the quantity of labour that was necessary to produce it.

It so happened that three economists, working independently of each other, came to use the concept of *utility* in order to arrive at a theory of how prices are determined. They were Stanley Jevons (English, 1835–1882), Carl Menger (Austrian, 1840–1921) and Léon Walras (French, 1834–1910). Although they differed widely in the methods they used, they all agreed on two fundamental points:

- the concept of *utility*, or satisfaction or pleasure derived from consuming something, is central to an idea of value that helps determine prices, and
- what matters is not the total utility of consuming something but rather the *extra* or *additional* utility of consuming one more unit of the good, known as *marginal utility*.

Suppose that utility can be measured in units of *utils* (imaginary units that measure satisfaction). According to Amandla's tastes and preferences, eating one ice cream provides her with 5 utils of satisfaction. If she eats a second one, she receives a total utility for the two ice creams of 9 utils; her total utility has increased by only 4 utils because she enjoyed the second ice cream less than the first. Hence, her marginal utility is 4 utils. As she enjoys each successive ice cream less than the previous one, her marginal utility keeps falling.

Some years later, in the early part of the 20th century, Alfred Marshall (English, 1842–1924) used these ideas to come up with the law of demand and the demand curve that we are familiar with today. If Amandla gets less and less marginal utility from consuming more ice creams, she will only be willing to buy more if their price falls, hence the law of demand. The entire analysis of demand and supply and market equilibrium that we study today is attributable to Alfred Marshall.

The concept of *marginal* is very important in economics. It will be encountered several times (Chapters 2 and 7, at HL only).

In the 19th century, the concept of **utility**, underlying utilitarianism, referring to the satisfaction derived from consuming something, was combined with the concept of **marginal**, meaning extra or additional, leading eventually to **marginal utility** as the basis of a theory of value that determines prices of goods and services. It forms the basis of rational consumer behaviour that is used to the present day in microeconomics.

Say's law in classical macroeconomics

Whereas we refer here to classical microeconomics and macroeconomics, these terms did not exist at the time, as there was no distinction during the 19th century between the micro and macro levels of analysis. As we will see below, macroeconomics as a distinct branch of economics was born in the 20th century.

The classical economists of the 19th century believed that the problem of unemployment could not arise under normal circumstances. While there could be occasional disturbances in an economy due to such events as wars, droughts or other major disruptions, ordinarily unemployment could not arise for extended periods because it was thought that the economy would keep on producing as much as is required in order to keep workers fully employed.

This idea came to be known as *Say's Law*, after the French economist Jean-Baptiste Say (1767–1832). Say was strongly influenced by Adam Smith, and was an advocate of free markets and laissez faire. The law that was named after him stated very simply that *supply creates its own demand*. What this means is that overall spending in an economy cannot fall enough to prevent all the output produced from being bought.

We can see what this means very clearly by examining the circular flow of income model presented in Section 1.3. Firms produce goods and services, and they pay households for providing them with the resources. The households receive this income and spend it to buy the goods and services. Therefore *supply creates its* *own demand*; the output that firms produce provides households with the income they need to buy that output. Therefore, workers will keep on working to produce that output and there will be full employment.

This simple idea came to be very seriously questioned during the Great Depression of the 1930s, which saw falling output and very high unemployment rates in many countries. These events gave rise to the birth of macroeconomics as a distinct branch of economics and the development of new theories to explain unemployment.

According to **Say's Law**, supply creates its own demand, a theory that claims that the economy tends toward full employment in the absence of any government intervention.

The Marxist critique of classical economics

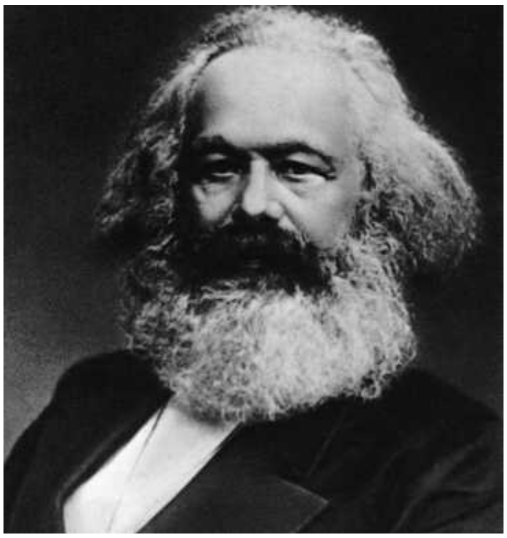


Figure 1.8: Karl Marx, German philosopher and political economist, whose theories formed the basis of modern communism, author of *Capital: Critique of Political Economy* (*Das Kapital* for short)

Karl Marx (1818–1883) was a German philosopher who had a profound influence on the course of global events during the 20th century, as many of his ideas formed the ideological basis for the establishment of communist regimes around the world. In addition to being a philosopher, Marx was an economist, historian, political theorist, sociologist and linguist.

Marx produced a huge amount of written work but perhaps the best known is his book *Capital: Critique of Political Economy* (for short *Das Kapital* in German, 1867), which was written over eighteen years. In this book, Marx presented what he considered to be a scientific account of the laws of capitalist development, which would eventually result in the downfall of capitalism. (*Capitalism* is the free market system, based on private ownership of the means of production, and driven by the desire to make profits.)

Marx's analysis is based on a version of the labour theory of value, which as noted earlier was the standard explanation of value in the 19th century, until the appearance of marginal analysis. The value of a good was determined by the amount of labour that was used to produce it. However, Marx observed that the price actually paid for a good was usually far greater than the value of labour that was put into producing it, measured by the wages paid to labour. The difference between the two, or the price of the good minus the value of labour to produce it, was termed *surplus value* by Marx, and corresponded to profit made by the owners of the factories.

Surplus value represented exploitation of workers by the owners of factories. The factory owners had an incentive to pay workers as little as possible, forcing them to work under appalling conditions, so that they could make the surplus value and hence their profits as large as possible. It is beyond doubt that conditions in 19th century factories in England, where Marx was living at the time, were truly appalling, as the writings of authors such as Charles Dickens clearly show.

As a historian, Marx saw economic systems being transformed over time in a particular order. Feudalism had been replaced by capitalism, and according to the historical laws he claimed to have discovered, capitalism was going to be replaced in the future by communism. This would happen because of the innate instability of capitalism. Competition would force capitalists (the owners of factories) to keep investing in new machinery in order to reduce labour costs and hence beat their competitors. But by using machines in place of labour, the capitalists would reduce the surplus value of labour that gave rise to their profits. Therefore the result would be declining profits for the capitalists.

At the same time that capitalists' profits were expected to fall, workers would become increasingly unemployed and poverty-stricken as they were replaced by more machines. Workers were expected to become more and more pitted against capitalists. The capitalist system would eventually be overthrown, to be replaced by communism where the means of production (the factories) would be owned collectively by the people.

Marx did not at any time describe his visions of communism in a systematic way, but it is clear from his writings that the systems that came into being in the 20th century were not based on his ideas. His predictions have not materialised, since capitalist profits do not appear to be falling. While capitalism does periodically undergo crises, such as the Great Depression of the 1930s and the global financial crisis of 2008, it does not appear to be on the verge of collapse. The communist systems that came into being in the 20th century were not the result of capitalism's collapse, but rather the result of force.

Yet Marx continues to be highly influential because he had keen insights into the workings of capitalism which are still highly relevant today. These include the recurrence of crises, which is similar to what is known today as the business cycle, the impoverishment of the middle class as income inequalities grow rapidly, the lack of growth in real wages over decades in many developed countries (related to the growth in income inequalities), increasing job insecurities and the risks of growing unemployment due to rapid technological change. Moreover, Marx has had a profound influence in other social sciences, including sociology, anthropology and political science.

Karl Marx is still widely read today, and there has been a growing interest in his work since the onset of the global financial crisis in 2008. Interestingly, a search on Amazon UK yields more results for Karl Marx than for Adam Smith.

Karl Marx developed a theory according to which *capitalism would be eventually replaced by communism* because of the market system's internal contradictions that would lead to its collapse. While this has not materialised, Marx is still highly regarded for his numerous insights into how capitalism works.

The 20th century

The Keynesian revolution

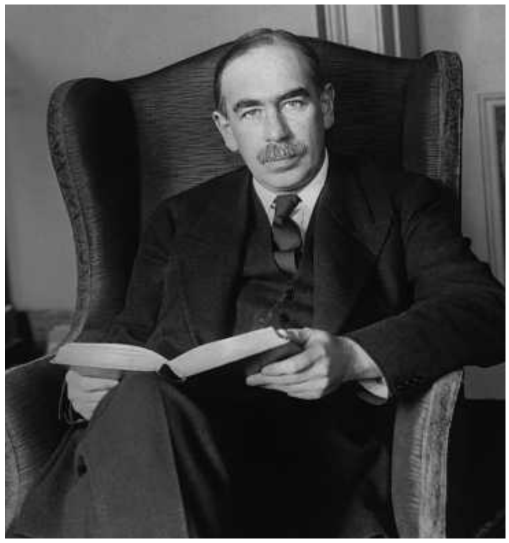


Figure 1.9: John Maynard Keynes, British economist, whose theories replaced classical economics, author of The General Theory of Employment, Interest and Money (*The General Theory* for short)

It was noted earlier that in the 19th century, it was believed that, in accordance with Say's Law, supply creates its own demand. In this way of thinking, it was not possible to have extended periods of unemployment. Therefore when the Great Depression of the 1930s occurred, bringing with it very significant declines in output along with high unemployment in many countries, economists were at a loss to explain how this had occurred.

John Maynard Keynes (1883–1946), an extremely influential British economist, set forth an economic theory that entirely replaced the classical theory and Say's Law. Keynes is best known for his book *The General Theory of Employment, Interest and Money* (for short, *The General Theory*, 1936).

Keynes argued that the state of full employment in an economy was only a special case that could not occur all the time. If spending decreased, there was nothing to ensure that an economy would return to a situation of full employment on its own. Classical economists thought that if overall demand decreased and there was a fall in spending, then prices will fall, which will cause spending to increase again, so output and employment will once again go to the full employment level. Also, classical economists thought that if there was a decrease in spending so that output fell, then there would be some temporary unemployment that would cause wages to fall, making employers increase the number of workers they hire. This would also bring employment back up to the full employment level once again.

But according to Keynes, wages were 'sticky', meaning they could not fall easily. In addition, falling wages meant that workers would have less money to spend, which would cause overall demand and spending to fall further. But sticky wages meant sticky prices, because producers could not lower their prices if they had to go on paying the same levels of wages. As a result, the economic system could not on its own go back to full employment.

Keynes argued that a situation like this requires government intervention in the form of increased government spending, which would give the economy the push it needed to get it going again. Using the idea of the *multiplier*, he showed that if the government increased its own spending on such things as building roads or schools, there would be a multiplied spending effect in the economy. This would help the economy to get out of its state of low output and unemployment.

Keynes thoroughly dismissed the laissez faire economy of the 19th century and provided a justification for government intervention that is urgently needed to avoid prolonged recessions (negative growth) and high unemployment. We will discuss Keynes' contribution in Chapter 9.

> One of the important contributions of John Maynard Keynes is the idea that *an economy left on its own will not necessarily lead to full employment*, thus requiring government intervention in order to ensure that full employment will be achieved.

The emergence of macroeconomic policy

Macroeconomic policy was inspired by the work of John Maynard Keynes discussed above. The term macroeconomics, as a field of economics distinct from microeconomics, did not even exist until 1945 when it was coined by Jacob Marschak.[6] According to the thinking of the classical economists, there was no need for macroeconomic policy since the economy was assumed to correct itself automatically through market forces in the absence of government intervention. We will discuss macroeconomic policies in detail in Chapter 13.

The monetarist/new classical counter-revolution

In the early 1970s, the global economy experienced cost-push inflation, also known as stagflation, on account of the first oil price crisis. Keynesian economics, with its focus on aggregate demand, was unable to provide a solution to inflation of this type. These events paved the way for the emergence and growing popularity of two schools of thought: **monetarism** and **new classical economics**.

Monetarism, attributed to the Nobel Prize winning economist Milton Friedman (1912–2006), emphasises the role of money in the economy. It is argued that changes in the money supply have major effects on output in the short run, and on the price level in the long run. New classical economics, associated partly with another Nobel Prize winning economist Robert Lucas (born 1937) emphasises the importance of individuals' 'rational expectations' of inflation and government policy actions.

While these two theories are quite different from each other, we are considering them together because they share a unifying principle regarding *the role of markets in bringing the economy back to a situation where there is full employment without any government intervention*. In fact, according to these two schools of thought, it is government intervention itself, in such areas as minimum wage laws and the operation of trade unions, that leads to sticky wages that do not fall, thus preventing the automatic adjustment of the economy back to full employment. If all wages and all prices could respond freely to the forces of demand and supply, markets on their own would achieve high levels of output and full employment.

It is clear from the above that this approach involves a rejection of Keynesian economics, which requires government intervention for the smooth functioning of an economy. It advocates a return to the classical idea of automatic full employment through the workings of a laissez-faire economy; hence the use of the term *new classical*.

[6] https://www.economist.com/briefing/2009/07/16/the-other-worldly-philosophers

The monetarist/new classical approach has been the inspiration of market-based supply-side policies. We will discuss this model and its implications for policy in Chapters 9 and 13.

According to the monetarist/new classical schools of thought, government intervention prevents the economy from reaching a state of full employment on its own, whereas a free market economy without the government intervening will tend toward full employment.

The 21st century

Behavioural economics and the dialogue with psychology

Since the early part of the 21st century, there has been a growing dialogue between economists and psychologists, resulting in a new field of economics known as **behavioural economics**. Behavioural economics questions the idea that marginal utility underlies demand and rational consumer behaviour (which we discussed earlier). It is argued that consumers do not have the necessary information available, but also and most importantly the human mind works in ways that are not rational in the ways the theory presupposes.

Rather than rely on a theory that explains the behaviour of consumers, behavioural economics relies on experiments and the accumulation of evidence about how consumers behave under a broad variety of different circumstances. This information is then used to formulate economic policies that will encourage consumers to behave in ways that are held to be socially desirable.

The growing importance and influence of behavioural economics is evidenced by the fact that, in a span of 15 years, three Nobel Prizes have been awarded to scholars dealing with this subject. They include Daniel Kahneman (2002) for his work on human judgement and decision-making under uncertainty; Robert J Shiller (2013) for his analysis of asset prices in the area of behavioural finance; and Richard Thaler (2017) for his work showing that people are predictably irrational in ways that contradict economic theory. (Behavioural economics will be studied at HL in Chapter 2.)

The growing awareness of the interdependence between the economy, society and the environment, and the need to move toward a circular economy

Earlier in this chapter we discussed the issue of sustainability, which arises from conditions of scarcity of resources. We learned that sustainability refers to the maintenance of resource quantities and quality over time. Until recently, economists carried out their work in isolation, without taking into consideration the interrelationships and interdependencies that exist between the economy, society and the environment. In recent years there is increasing awareness of the need to consider three pillars of sustainability together: the economy, society and the environment, sometimes abbreviated as profit, people, planet.

In addition, it is imperative to move away from the traditional extractive take–make–dispose model, which characterises our current approach to producing products and disposing of waste. This approach involves extracting resources, making them into products and then throwing them out. The imperative of achieving sustainable development requires that we adopt a **circular economy**. The idea behind a circular economy is that goods should be produced in such a way that they can be repaired rather than thrown out. In addition, they would be made out of biological materials so that once discarded they can go back to the biosphere and prevent pollution of the planet. According to London's Waste and Recycling Board:

> *'A circular economy is a more efficient and environmentally sound alternative to the traditional linear economy. It is one in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end their life.'*[7]

We will study the problem of sustainable resource use in Chapter 5.

[7] https://www.lwarb.gov.uk/

Novel thinking in the 21st century includes: (i) the contributions of psychology to economics, which offer alternative ways of understanding how consumers make decisions, with a view to influencing consumer choices toward socially desirable outcomes; and (ii) the development of new models regarding sustainability that focus on the close interdependencies between the economy, society and the environment, and the concept of a circular economy.

TEST YOUR UNDERSTANDING 1.14

1. Referring to the concept of laissez faire, explain Adam Smith's main contribution to economics.
2. Outline the three phases of macroeconomic thought on the need for government intervention in the economy for the purpose of achieving full employment, from the 19th century to the late 20th century.
3. Explain how perceptions regarding the concept of *utility* have evolved from the 19th century to the present.
4. Outline the meaning of a *circular economy* and its importance to sustainable development.
5. Select a famous economist you are interested in and present his/her work, either individually or in a group.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Unit 2
Microeconomics

Microeconomics is concerned with the behaviour of consumers, firms and resource owners, who are the most important economic decision-makers in a market economy. We will study the model of demand and supply, which forms the basis of the market economy and is one of the most important analytical tools in microeconomics. We will learn about the benefits of free markets as well as their imperfections in a variety of situations where they fail to meet important economic objectives. We will also examine the role of governments. We will see what effects governments have when they interfere in markets as well as how they can help achieve better social outcomes when markets fail to perform well.

The tools we will develop in microeconomics are important because they provide many insights into the workings of the market economy, and into the effects of different types of government intervention. In addition, they form the basis of additional topics we will study in later parts of this book.

Real world issue 1: How are choices made by consumers and producers when they try to meet their economic objectives?

CONCEPTUAL UNDERSTANDINGS

1. The economy's resources are allocated (assigned) to the production of goods and services intended to meet needs and wants mainly through the interactions of consumers and producers in markets.
2. The ***choices*** made by consumers and producers are the result of complex decision-making.
3. The achievement of allocative ***efficiency*** means that welfare is maximised (made the greatest it can be).
4. Continuous ***change*** gives rise to dynamic markets.

These topics are addressed in Chapters 2 and 3. In Chapter 2 we will discover the basic building blocks of microeconomics: demand, supply and markets. We will examine how consumers and producers interact in markets and we will see in more detail what Adam Smith meant by the *invisible hand*. We will also examine some recent developments that offer alternative explanations of consumer and firm behaviour.

Chapter 3 will introduce a new concept, that of *elasticity*, which measures how consumers and firms respond to changes in prices, and allows us to better understand the workings of markets.

A boy getting school supplies

Chapter 2

Competitive markets: Demand and supply

BEFORE YOU START

- Over time, you see prices of goods and services change. Why do you think sellers of goods and services change their prices?
- When you buy goods and services, do you think your choices are always in your best self-interest?
- How do consumers and producers make choices in trying to meet their economic objectives?

In this chapter we examine what lies at the heart of every market-based economy: the forces of demand and supply. We will then see how the interactions of demand and supply arrive at equilibrium market prices, and we will study the special qualities of free competitive markets. This chapter will also examine some recent developments in the fields of consumer and producer behaviour that provide novel explanations on how these fundamental decision-makers make choices.

Photo: Fresh vegetables for sale

2.1 Introduction to competitive markets

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- outline the meaning of a **competitive market** (AO1)

Markets

The nature of markets

A market originally was a place where people gathered to buy and sell goods. Such markets still exist today, for example cattle markets, fish markets, fruit and vegetable markets and flea markets, involving a physical meeting place where buyers and sellers meet face to face.

The term **market** has since evolved to include any kind of arrangement where buyers and sellers of goods, services or resources are linked together to carry out an exchange.

The market may be in a specific place (such as a vegetable market), or it may involve many different places (such as the oil market). Buyers and sellers may meet (say, in a shop), or they may never meet, communicating by fax, phone, internet, classified ads, or any other method which allows them to convey information about price, quantity and quality.

A market can be local, where the buyers and sellers originate from a local area; it may be national, in which case the buyers and sellers are from anywhere within a country; or it may be international, with buyers and sellers from anywhere in the world. For example, small neighbourhood bakeries produce and sell bread and other baked goods for the local community – this is a local market. Local takeaway restaurants also produce for the local market. The labour market, on the other hand, tends to be mostly a national market. By contrast, the world oil market includes oil producers in different countries, and buyers of oil virtually everywhere in the world, as well as wholesalers, retailers and other intermediaries involved in buying and selling oil around the world.

Goods and services are sold in product markets, while resources (factors of production) are sold in resource markets (factor markets).

TEST YOUR UNDERSTANDING 2.1

1 What is a market?

2 Suggest more examples of local, national and international markets.

The meaning of a competitive market

Competition is generally understood to be a process in which rivals compete in order to achieve some objective. For example, firms may compete with each other over who will sell the most output, consumers may compete over who will buy a scarce product, workers compete over who will get the best jobs with the highest salaries, countries compete over which will capture the biggest export markets.

Beyond this everyday sense, competition in microeconomics occurs when there are many buyers and sellers acting independently, so that no one has the ability to influence the price at which a product is sold. This should be contrasted with *market power* (also known as *monopoly power*) which refers to the control that a seller may have over the price of the product it sells. The greater the market power, the greater is the control over price. On the other hand, the greater the degree of competition between sellers, the smaller their market power, and the weaker is their control over the price.

In this chapter we will study **competitive markets** composed of large numbers of sellers and buyers acting independently, so that no one individual seller or small group of sellers has the ability to control the price of the product sold. Instead, the price of the product is determined by the interactions of many sellers and buyers, through the forces of demand and supply.

2.2 Demand

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the law of demand (AO2)
- draw a demand curve (AO4)
- explain the relationship between individual demand and market demand (AO2)
- analyse the non-price determinants of demand as causes of demand curve shifts (AO2)
- using diagrams distinguish between movements along the demand curve and shifts of the demand curve (AO4)
- explain the assumptions that underlie the law of demand (HL only) (AO2)
  - the law of diminishing marginal utility
  - the income and substitution effects

Understanding the law of demand and the demand curve

Demand is concerned with the behaviour of buyers. Consumers (or households) are buyers of goods and services in product markets, whereas firms (or businesses) are buyers of factors of production in resource markets. In our analysis of demand and supply we will focus mainly on product markets and therefore on the behaviour of consumers as buyers (though the same general principles described here apply also to the behaviour of firms as buyers in resource markets).

Individual demand

As buyers, consumers are demanders of those items they wish to buy.

The **demand** of an individual consumer indicates the various quantities of a good (or service) the consumer is *willing and able to buy* at different possible prices during a particular time period, *ceteris paribus*.

A consumer's demand for a good can be presented as a demand schedule, or as a table listing quantity demanded at various prices. Table 2.1 shows a consumer's demand schedule for chocolate bars. When the price of chocolate bars is $5, the consumer is willing and able to buy two chocolate bars in a week. When the price is $4, the consumer is willing and able to buy four chocolate bars in a week, and so on.

Price of chocolate bars ($)	Quantity of chocolate bars demanded (per week)
5	2
4	4
3	6
2	8
1	10

Table 2.1: Demand schedule for a consumer

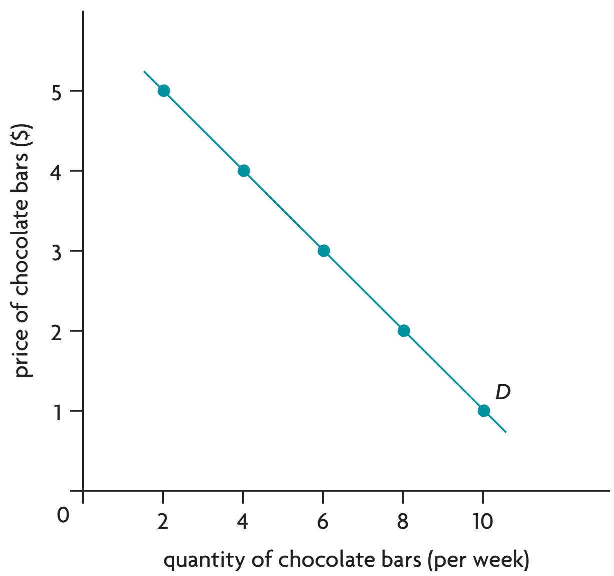


Figure 2.1: Demand curve for an individual consumer

'Willing' means the consumer wants to buy the good; 'able' means that the consumer can afford to buy it. (You may want to buy a Ferrari, but can you afford it? If not, your desire to buy one will not show up as demand for Ferraris. Or, you can afford a Ferrari but you have no desire to own one; again you will not have any demand for Ferraris.)

Ceteris paribus means that all things other than price that can affect how much the consumer is willing and able to buy are assumed to be constant and unchanging (see Chapter 1, and 'Quantitative techniques' chapter in the digital coursebook). In fact, the consumer's demand is affected not only by price, but also by many other things, like income, tastes and prices of related goods. For the moment, we ignore all these and concentrate only on the relationship between the quantity of a good and its price.

The information contained in the demand schedule can be plotted as a graph, shown in Figure 2.1. The price of chocolate bars is plotted on the vertical axis and quantity of chocolate bars on the horizontal axis. The curve in Figure 2.1 is a **demand curve**. Note that even though this is a straight line, it is referred to as a 'curve'.

The demand schedule and demand curve do not tell us anything about how many chocolate bars the consumer will actually buy and what price the consumer will pay. This information will be given to us later through the interaction of demand with supply. The demand information only tells us how many chocolate bars the consumer would be prepared to buy if the price were $5, or $4, and so on.

The law of demand

The demand curve plotted in Figure 2.1 illustrates a very important relationship: as the price of a good falls, the quantity of the good demanded increases. When two variables change in opposite directions, so that as one falls, the other increases, they are said to have a 'negative' (or 'indirect') relationship. This relationship is a 'causal' one, because changes in price *cause* changes in quantity demanded. The negative relationship between price and quantity demanded is known as the law of demand.

According to the **law of demand**, there is a negative relationship between the price of a good and its quantity demanded over a particular time period, *ceteris paribus:* as the price of the good increases, quantity demanded falls; as the price falls, quantity demanded increases, *ceteris paribus*.

The law of demand is most likely to be consistent with your experience. The higher the price of a good, the less of it you are probably willing and able to buy.

From individual demand to market demand

So far we have considered the demand for a good of one individual consumer. *Market demand* shows the total quantities in the market for the good consumers are willing and able to buy at different prices (during a particular period of time, *ceteris paribus*). Market demand is the sum of all individual demands for that good. Figure 2.2 shows how the quantity demanded by consumer A is added to the quantity demanded by consumer B, and so on until all the quantities demanded by all consumers of chocolate bars are added up. (Note that consumer A has a different demand for chocolate bars than consumer B, indicating different preferences). For example, at

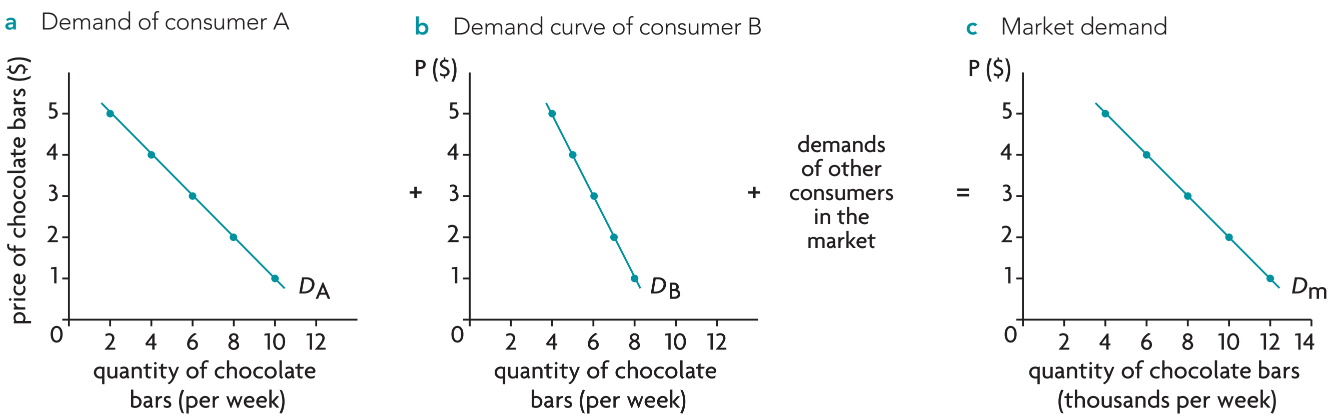


Figure 2.2: Market demand as the sum of individual demands

the price of \$4, we add the four bars demanded by consumer A to the five bars demanded by consumer B, and so on to all the quantities demanded by other consumers, to arrive at the sum of 6000 chocolate bars per week. This sum is a point on the market demand curve D_m. When we add individual demands in this way for each of the possible prices, we derive the entire market demand curve D_m, showing the total demand in the chocolate bar market.

> **Market demand** is the sum of all individual demands for a good. The market demand curve illustrates the law of demand, shown by the negative relationship between price and quantity demanded.

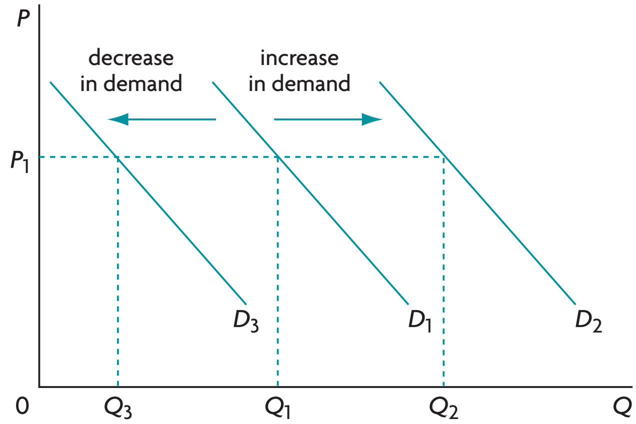


Figure 2.3: Shifts in the demand curve

Non-price determinants of demand and shifts of the demand curve

The non-price determinants

The **non-price determinants of demand** are the variables other than price that can influence demand. They are the variables assumed to be unchanging by use of the *ceteris paribus* assumption.

Changes in the non-price determinants of demand cause shifts in the demand curve: the entire demand curve moves to the right or to the left. In Figure 2.3, note that the vertical axis is labelled '*P*', standing for price, and the horizontal axis is labelled '*Q*', standing for quantity. Suppose the original demand curve is given by D_1. If price is P_1, then the demand curve D_1 indicates that quantity Q_1 will be demanded. If the demand curve shifts to the right, to D_2, at the same price P_1 a larger quantity, Q_2, will be demanded. If, on the other hand, the demand curve shifts to the left, from D_1 to D_3, then a smaller quantity, Q_3, will be demanded at the same price P_1.

A rightward shift of the demand curve indicates that more is demanded for a given price; a leftward shift of the demand curve indicates that less is demanded for a given price. A rightward shift of the curve is called an *increase in demand*; a leftward shift is called a *decrease in demand*.

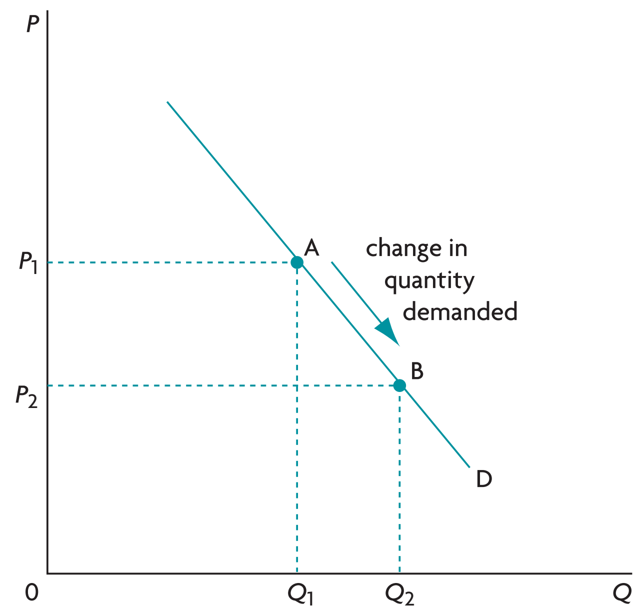
The non-price determinants of market demand include:

- **Income in the case of normal goods.** A good is a **normal good** when demand for it increases in response to an increase in consumer income (demand for the good varies directly with income). Most goods are normal goods. Therefore, an increase in income leads to a rightward shift in the demand curve, and a decrease in income leads to a leftward shift.
- **Income in the case of inferior goods.** While most goods are normal, there are some goods where the demand falls as consumer income increases; the good is then an **inferior good** (the demand for the good varies inversely with income). Examples of inferior goods are second-hand clothes, used cars and bus tickets. As income increases, consumers switch to more expensive alternatives (new clothes, new cars and cars or aeroplanes rather than travelling by bus), and so the demand for the inferior goods falls. Thus an increase in income leads to a leftward shift in the demand curve and a decrease in income produces a rightward shift.
- **Preferences and tastes.** If preferences and tastes change in favour of a product (the good becomes more popular), demand increases and the demand curve shifts to the right; if tastes change against the product (it becomes less popular), demand decreases and the demand curve shifts to the left.
- **Prices of substitute goods.** Two goods are **substitutes (substitute goods)** if they satisfy a similar need. An example of substitute goods is Coca-Cola® and Pepsi®. A fall in the price of one (say, Coca-Cola) results in a fall in the

demand for the other (Pepsi). The reason is that as the price of Coca-Cola falls, some consumers switch from Pepsi to Coca-Cola, and the demand for Pepsi falls. On the other hand, if there is an increase in the price of Coca-Cola, this will result in an increase in the demand for Pepsi as some consumers switch from Coca-Cola to Pepsi. Therefore, for any two substitute goods *X* and *Y*, a decrease in the price of *X* produces a leftward shift in the demand for *Y*, while an increase in the price of *X* produces a rightward shift in the demand for *Y*. In brief, in the case of substitute goods, the price of *X* and demand for *Y* change in the same direction (they both increase or they both decrease). Other examples of substitute goods are oranges and apples, Cadbury's and Nestlé chocolate, and milk and yoghurt.

- **Prices of complementary goods.** Two goods are **complements (complementary goods)** if they tend to be used together. An example of complementary goods is computers and computer software. In this case, a fall in the price of one (say, computers) leads to an increase in the demand for the other (computer software). This is because the fall in the price of computers results in a bigger quantity of computers being purchased, and the demand for computer software increases. Therefore, for any two complementary goods *X* and *Y*, a fall in the price of *X* leads to a rightward shift in the demand for *Y*, and an increase in the price of *X* leads to a leftward shift in the demand for *Y*. In the case of complementary goods, the price of *X* and the demand for *Y* change in opposite directions (as one increases, the other decreases). More examples of complementary goods are shoes and shoe laces, tennis shoes and tennis rackets, and table-tennis balls and table-tennis rackets. Note that most goods are not related to each other; these are called independent goods. For example, pencils and apples, cars and ice cream, telephones and books are unrelated to one another, and the change in the price of one will have little or no effect on the demand for the other.
- **The number of consumers.** If there is an increase in the number of consumers (demanders), demand increases and therefore the market demand curve shifts to the right; if the number of consumers decreases, demand decreases and the curve shifts to the left. This follows simply from the fact that market demand is the sum of all individual demands.

a A movement along the demand curve, caused by a change in price, is called a 'change in quantity demanded'



b A shift of a demand curve, caused by a change in a determinant of demand, is called a 'change in demand'

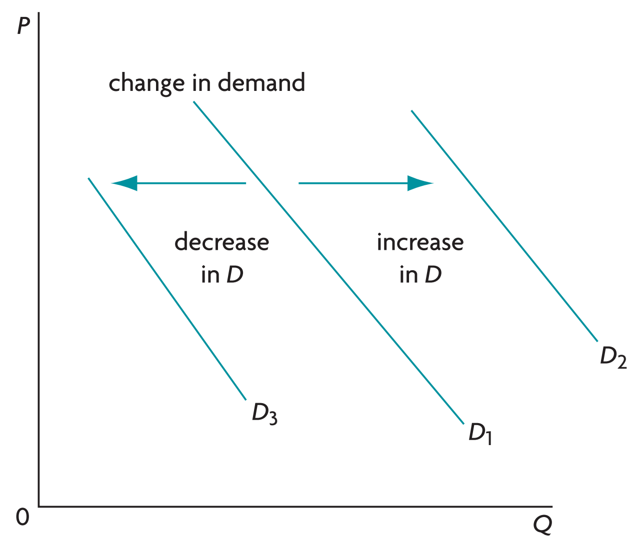


Figure 2.4: Movements along and shifts of the demand curve

Movements along a demand curve and shifts of the demand curve

It is important to distinguish between movements along a demand curve, and shifts of a demand curve. Whenever the price of a good changes, *ceteris paribus*, it leads to a movement along the demand curve. In Figure 2.4(a), if the price falls from P_1 to P_2, the quantity of the good demanded increases from Q_1 to Q_2. A movement along the demand curve from A to B has occurred; this is referred to as an *increase in quantity demanded*. An increase in price gives rise to a *decrease in quantity demanded*, resulting in a movement from B to A.

By contrast, any change in a non-price determinant of demand results in a shift in the entire demand curve, as shown in Figure 2.4(b); this is called a *change in demand*. For example, if there is an increase in the number of buyers, the demand curve shifts rightward from D_1 to D_2; this is called an *increase in demand*, shown in Figure 2.4(b). A decrease in the number of buyers causes a leftward shift of the demand curve from D_1 to D_3; this is called a *decrease in demand*. To summarise:

> Any change in price produces a *change in quantity demanded*, shown as a movement on the demand curve. Any change in a non-price determinant of demand leads to a *change in demand*, represented by a shift of the entire demand curve.

TEST YOUR UNDERSTANDING 2.2

1 a Define 'demand'.

b State the law of demand.

c Explain whether the law of demand shows a negative or positive relationship.

d Show the law of demand in a diagram.

e Describe the relationship between individual demand and market demand.

f Distinguish between a 'change in demand' and a 'change in quantity demanded' and explain the cause or causes of each.

CONTINUED

g How would you show the difference between a movement along the demand curve and a shift of the demand curve in a diagram?

h Identify the non-price determinants of demand.

2 Using diagrams, show the impact of each of the following on the demand curve for product A.

a The number of consumers in the market for product A increases.

b Consumer income increases and product A is an inferior good.

c Consumer income decreases and product A is a normal good.

d A news report claims that use of product A has harmful effects on health.

e The price of substitute good B falls.

f The price of complementary good B increases.

Assumptions underlying the law of demand (HL only)

The law of diminishing marginal utility

The law of *diminishing marginal utility* is based on a simple theory of consumer behaviour that explains the negative relationship between price and quantity demanded, or the law of demand.

Utility is the satisfaction that consumers gain from consuming something. It is a subjective concept, because satisfaction is something that depends entirely on personal tastes and preferences, which vary from person to person.

Utility cannot be measured, but for the purposes of developing the theory, we assume that utility is quantifiable (it can be measured in terms of numbers of units), and the unit that we use is the *util*. Utils do not exist in actual fact, however we assume they exist for the purposes of building the theory.

Total utility is the total satisfaction that consumers get from consuming something, while *marginal utility* is the extra satisfaction that consumers receive from consuming one more unit of a good. Table 2.2 shows how total utility and marginal utility are related to each other. You can see that both total utility and marginal utility are measured in utils.

Table 2.2 shows Anna's preferences when she buys T shirts. The first T shirt provides her with 15 utils of total utility. This is also her marginal utility. When she buys a second T shirt, her total utility increases from 15 utils to 27 utils, but her marginal utility, or extra satisfaction from buying the second T shirt has fallen to 12 utils. The marginal utility of the second unit is simply the total utility of the second unit minus the total utility of the first unit, which is 27 utils – 15 utils = 12 utils. And so on with each extra T shirt she buys. Her total utility increases, but her marginal utility falls. Marginal utility keeps decreasing and at some point it may even become negative, meaning that total utility starts to fall. This pattern illustrates the law of diminishing marginal utility, which is based on the idea that the satisfaction consumers get from consuming more and more units of a good decreases.

The law of diminishing marginal utility explains the law of demand. As we know, according to the law of demand, as the price of a good decreases, quantity demanded increases *ceteris paribus*. The law of diminishing marginal utility shows that if the consumer derives less and less utility from each extra unit of a good consumed, then she or he will buy additional units only if the price of the good falls.

According to the **law of diminishing marginal utility**, as consumption of a good increases, marginal utility, or the extra utility the consumer receives, decreases with each additional unit consumed. This underlies the law of demand, as it shows that a consumer will be willing to buy an additional unit of a good only if its price falls.

Number of T shirts bought per year	Total utility (Utils)	Marginal utility (Utils)
0	0	–
1	15	15
2	27	12
3	36	9
4	42	6
5	45	3
6	45	0
7	42	–3

Table 2.2: Total and marginal utility

The concept of marginal utility is the basic building block of the theory of consumer behaviour, according to which consumers arrange their purchases of many different goods in a way that will allow them to *maximise the total utility* they derive.

The income and substitution effects

The income and substitution effects are an alternative explanation of the law of demand.

As we have seen, the law of demand shows the relationship between the price of a good and quantity of the good demanded, *ceteris paribus*. Any price change causes a change in quantity demanded, shown as a movement along the demand curve. We will now see that the total effect of a price change on quantity demanded can be broken down into two separate effects: the substitution effect and the income effect, with the total effect of a price change being the sum of these two effects.

- The **substitution effect**: If the price of a good falls, the consumer substitutes (buys more) of the now less expensive good. Therefore quantity demanded increases. There is always a negative relationship between price and quantity demanded as a result of the substitution effect: as price falls, the quantity of the good demanded increases; as price increases, quantity demanded falls.

- The **income effect**: Consider again a fall in price. This means that the consumer's *real income* (or purchasing power) has increased. (To understand what this means consider the following example. Say you have $12 and you want to buy some pencils. When the price is $4 per pencil, you can buy three pencils. Suppose then that the price of pencils falls to $3 per pencil. You can now buy four pencils. The sum of money at your disposal ($12) has not changed, and yet the 'purchasing power' of $12, or what your money can buy, has increased as a result of the fall in the price of pencils. 'Real income' is the same as 'purchasing power'; it increases as prices fall, and it decreases as prices rise.) Therefore as price falls and real income increases, quantity demanded of the good increases. Once again there is a negative relationship between price and quantity demanded.

The substitution effect and the income effect reinforce each other: a fall in price leads to an increase in quantity demanded, both because of the substitution effect and because of the income effect.

In the case of most goods, whose price is a small fraction of income, the substitution effect is the most important part of the explanation of the law of demand, because being a small fraction of income, the effect that a price change has on real income will be tiny. The income effect becomes important only in the case of purchases of goods whose price take up a large fraction of income.

Note that the above analysis refers to *normal goods*, where income and demand change in the same direction; as real income increases, consumers want to buy more of the good. If you are interested in seeing what happens in the case of *inferior goods* you can read about it in the digital coursebook as Supplementary material.

TEST YOUR UNDERSTANDING 2.3

1 a Using an example, explain the law of diminishing marginal utility.

 b How is this law related to the downward sloping demand curve?

2 Explain how the income and substitution effects can explain the downward sloping demand curve.

2.3 Supply

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the law of supply (AO2)
- draw a supply curve (AO4)
- explain the relationship between individual supply and market supply (AO2)
- analyse the non-price determinants of supply as causes of supply curve shifts (AO2)
- using diagrams, distinguish between movements along the supply curve and shifts of the supply curve (AO4)
- explain the assumptions that underlie the law of supply (HL only) (AO2)
    - law of diminishing marginal returns
    - increasing marginal cost

Understanding the law of supply and the supply curve

Supply is concerned with the behaviour of sellers, which include firms in the product markets and households in resource markets. As we are focusing on product markets, we will consider the behaviour of firms as sellers (though the same general principles also apply to sellers of factors of production in resource markets).

Individual supply

Firms produce goods and services, and they supply them to product markets for sale. As sellers, therefore, they are suppliers of goods and services.

The **supply** of an individual firm indicates the various quantities of a good (or service) a firm is *willing and able* to produce and supply to the market for sale at different possible prices, during a particular time period, *ceteris paribus*.

A firm's supply of a good can be presented as a supply schedule, or a table showing the various quantities of a good the firm is willing and able to produce and supply at various prices. Table 2.3 shows a firm's supply schedule for chocolate bars. The same information appears as a graph in Figure 2.5, where price is plotted on the vertical axis and quantity on the horizontal axis. The line appearing in the diagram is the **supply curve** of the firm. If the price is $4, the firm supplies 500 chocolate bars in the course of a week; if price were $3, then the firm would supply 400 chocolate bars, and so on.

As in the case of demand, where price is only one thing that determines how much is demanded, so in the case of supply, price is only one thing that influences how much the firm supplies to the market; hence the *ceteris paribus* assumption. For the moment, we will ignore other possible influences on supply and focus only on the relationship between price and quantity.

Price of chocolate bars ($)	Quantity of chocolate bars supplied (per week)
5	600
4	500
3	400
2	300
1	200

Table 2.3: Supply schedule for a firm

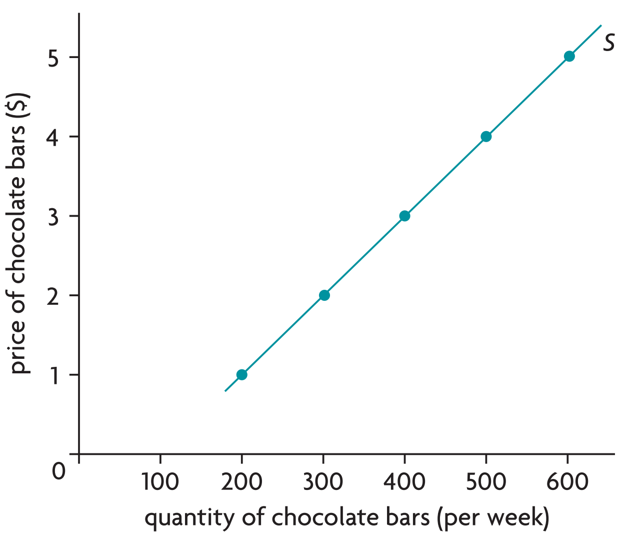


Figure 2.5: Supply curve for a firm

The supply schedule and the supply curve do not tell us anything about how many chocolate bars the firm will actually supply to the market nor what price the firm will receive. The supply information tells us only how many chocolate bars the firm would be prepared to produce and sell if the price were $5, or $4, and so on.

The law of supply

The supply curve in Figure 2.5 illustrates an important relationship: as price increases, quantity supplied also increases. When two variables change in the same direction (as one increases, the other also increases), they are said to have a 'positive' (or 'direct') relationship. This relationship is a 'causal' one, because changes in price *cause* changes in quantity supplied. The positive causal relationship between the two variables, price and quantity supplied, is summarised in the law of supply.

According to the **law of supply**, there is a positive relationship between the quantity of a good supplied over a particular time period and its price, *ceteris paribus*: as the price of the good increases, the quantity of the good supplied also increases; as the price falls, the quantity supplied also falls, *ceteris paribus*.

From individual supply to market supply

Market supply indicates the total quantities of a good that firms are willing and able to supply in the market at different possible prices and is given by the sum of all individual supplies of that good. Figure 2.6 provides an example where at each price, the quantity supplied by firm A is added to the quantity supplied by firm B, and so on, until all the quantities supplied by all firms producing chocolate bars are added up. For example, at the price of $3, firm A supplies 400 bars per week and firm B supplies 300 bars. If we add these quantities together with all the quantities supplied by other firms, we obtain 8000 bars per week, which is a point on the market supply curve, S_m, corresponding to the price of $3. When the firms' supplies are added up this way for each possible price, we derive the market supply curve, S_m.

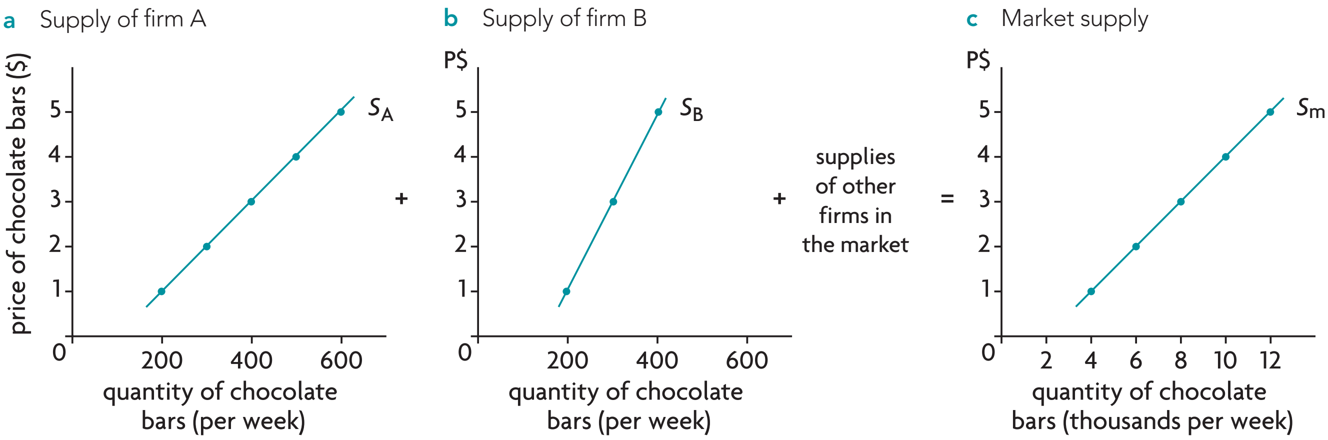


Figure 2.6: Market supply as the sum of individual supplies

Market supply is the sum of all individual firms' supplies for a good. The market supply curve illustrates the law of supply, shown by a positive relationship between price and quantity supplied.

The vertical supply curve

Under certain special circumstances, the supply curve is vertical at some particular fixed quantity, as in Figure 2.7. A vertical supply curve tells us that even as price increases, the quantity supplied cannot increase; it remains constant. The quantity supplied is independent of price. There are two reasons why this may occur:

- There is a fixed quantity of the good supplied because there is no time to produce more of it. For example, there is a fixed quantity of tickets in a theatre, because there is a fixed number of seats. No matter how high the price, it is not possible to increase the number of seats in a short period of time.
- There is a fixed quantity of the good because there is no possibility of ever producing more of it. This is the case with original antiques (for example, Stradivarius violins) and original paintings and sculptures of famous artists. It may be possible to make reproductions, but it is not possible to make more originals.

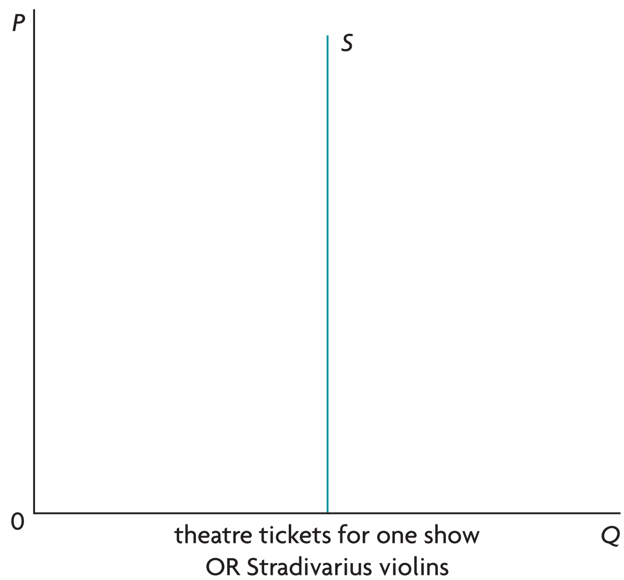


Figure 2.7: The vertical supply curve

Non-price determinants of supply and shifts of the supply curve

The non-price determinants

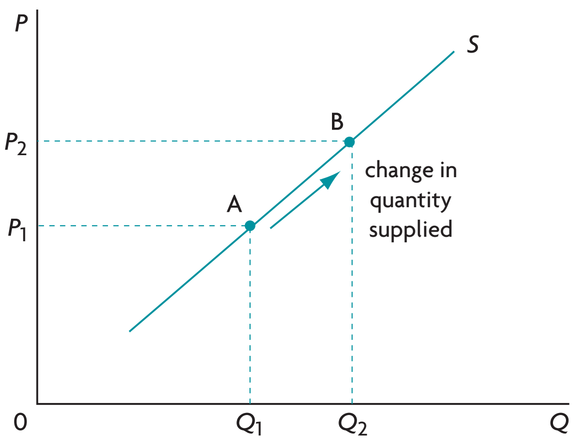
We now turn to the **non-price determinants of supply**, or the factors other than price that can influence supply. Changes in the determinants of supply cause shifts in the supply curve. A rightward shift means that for a given price, supply increases and more is supplied; a leftward shift means that for any price, supply decreases and less is supplied. As Figure 2.8(b) shows, when supply is S_1, quantity Q_1 will be supplied at price P_1. If there is an increase in supply to S_2, at the same price P_1, then Q_2 quantity is supplied. If supply falls to S_3, then Q_3 quantity is supplied at the same price P_1.

A rightward shift of the supply curve indicates that more is supplied for a given price; a leftward shift of the supply curve indicates that less is supplied for a given price. A rightward shift of the curve is called an *increase in supply*; a leftward shift is called a *decrease in supply*.

The non-price determinants of market supply include the following:

- **Costs of factors of production (factor or resource prices).** The firm buys various factors of production (land, labour, capital, entrepreneurship) that it uses to produce its product. Prices of factors of production (such as wages, which are the price of labour) determine the firm's costs of production. If a factor price rises, production costs increase, production becomes less profitable and the firm produces less; the supply curve shifts to the left. If a factor price falls, costs of production fall, production becomes more profitable and the firm produces more; the supply curve shifts to the right.
- **Technology.** A new improved technology lowers costs of production, thus making production more profitable. Supply increases and the supply curve shifts to the right. In the (less likely) event that a firm uses a less productive technology, costs of production increase and the supply curve shifts leftward.
- **Prices of related goods: competitive supply.** **Competitive supply** of two or more products refers to production of one or the other by a firm; the goods compete for the use of the same resources, and producing more of one means producing less of the other. For example, a farmer, who can grow wheat or corn, chooses to grow wheat. If the price of corn increases, the farmer may switch to corn production as this is now more profitable, resulting in a fall in wheat supply and a leftward shift of the supply curve. A fall in the price of corn results in an increase in wheat supply and a rightward shift of the supply curve.
- **Prices of related goods: joint supply.** **Joint supply** of two or more products refers to production of goods that are derived from a single product, so that it is not possible to produce more of one without producing more of the other. For example, butter and skimmed milk are both produced from whole milk; petrol (gasoline), diesel oil and heating oil are all produced from crude oil. This means that an increase in the price of one leads to an increase in its quantity supplied and also to an increase in supply of the other joint product(s).

a A movement along the supply curve, caused by a change in price, is called a 'change in quantity supplied'



b A shift of the supply curve, caused by a change in a determinant of supply, is called a 'change in supply'

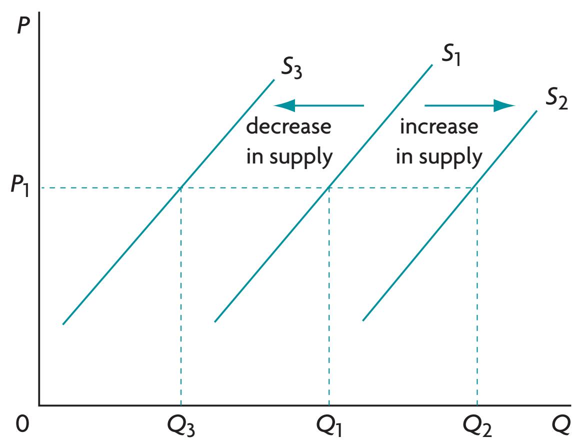


Figure 2.8: Movements along and shifts of the supply curve

- **Producer (firm) price expectations.** If firms expect the price of their product to rise, they may withhold some of their current supply from the market (not offer it for sale), expecting that they will be able to sell it at the higher price in the future; in this case, there is a fall in supply in the present and a leftward shift in the supply curve. If the expectation is that the price of their product will fall, supply increases in the present to take advantage of the current higher price, hence a rightward shift in the supply curve.
- **Taxes (indirect taxes or taxes on profits).** Firms treat taxes as if they were costs of production. Therefore, the imposition of a new tax or the increase of an existing tax represents an increase in production costs, so supply will decrease and the supply curve shifts to the left. The elimination of a tax or a decrease in an existing tax represents a fall in production costs; supply increases and the supply curve shifts to the right. (See Chapter 4.)
- **Subsidies.** A **subsidy** is a payment made to the firm by the government, and so has the opposite effect of a tax. (Subsidies may be given in order to increase the incomes of producers or to encourage an increase in the production of the good produced.) The introduction of a subsidy or an increase in an existing subsidy is equivalent to a fall in production costs, resulting in a rightward shift in the supply curve, while the elimination of a subsidy or a decrease in a subsidy leads to a leftward shift. (See Chapter 4).
- **The number of firms.** An increase in the number of firms producing the good increases supply, resulting in a rightward shift in the supply curve; a decrease in the number of firms decreases supply and produces a leftward shift. This follows from the fact that market supply is the sum of all individual supplies.
- **'Shocks', or sudden unpredictable events.** Sudden, unpredictable events called 'shocks', can affect supply, such as weather conditions in the case of agricultural products, war, or natural/man-made catastrophes. For example, the Louisiana oil spill in 2010 resulted in a decrease in the supply of locally produced seafood.

Movement along a supply curve and shift of the supply curve

Just as in the case of the demand curve, so in the case of the supply curve we distinguish between movements along and shifts of the entire curve. Movements along a supply curve can be caused only by changes in price. In Figure 2.8(a), as price increases from P_1 to P_2, quantity supplied increases from Q_1 to Q_2. There has been a movement along the supply curve from A to B. This is called a *change in quantity supplied*. If there is a change in a non-price determinant of supply, supply will increase or decrease, and the entire curve will shift to the right or to the left, as in Figure 2.8(b). This is called a *change in supply*.

> Any change in price produces a *change in quantity supplied*, shown as a movement on the supply curve. Any change in a determinant of supply (other than price) produces a *change in supply*, represented by a shift of the whole supply curve.

TEST YOUR UNDERSTANDING 2.4

1 a Define 'supply'.

b State the law of supply.

c Explain the relationship between price and the quantity supplied.

d Show the law of supply in a diagram.

e Describe the relationship between individual supply and market supply.

f Distinguish between a 'change in supply' and a 'change in quantity supplied' and explain the cause or causes of each.

g Draw two diagrams to show the difference between a movement along the supply curve and a shift of the supply curve.

h Identify the non-price determinants of supply.

CONTINUED

2 Give some examples of goods with a vertical supply curve.

3 Using diagrams, show the impact of each of the following on the supply curve of product A.

   a The number of firms in the industry producing product A decreases.

   b The price of oil, a key input in the production of product A, increases.

   c Firms expect that the price of product A will fall in the future.

   d The government grants a subsidy on each unit of A produced.

   e The price of product B falls, and B is in competitive supply with A.

   f The price of product B increases, and B is in joint supply with A.

   g A new technology is adopted by firms in the industry producing A.

Assumptions underlying the law of supply (HL only)

The law of supply is based on the relationship between production and the costs of production. We begin this section by introducing some new concepts that are essential to understanding the relationship between production and costs.

Law of diminishing marginal returns

The short run and the long run in microeconomics

All firms use *inputs* (or resources, or factors of production) to produce output. The quantities of inputs needed to produce output is determined by a technical relationship between inputs and output. This technical relationship depends on a distinction between the *short run* and the *long run*:

- The **short run** is a time period during which at least one input is fixed and cannot be changed by the firm. When we say fixed, we mean it is unchanging in quantity and quality. For example, if a firm wants to increase output, it can hire more labour and increase materials, tools and equipment, but it cannot easily change the size of its buildings, factories and heavy machinery. The buildings, factories and heavy machinery that are unchanging are *fixed*, whereas the labour and materials are *variable*. As long as the firm has at least one fixed input, it is operating in the short run.
- The **long run** is a time period when all inputs can be changed. Using the example above, in this time period the firm can build new buildings and factories and buy more heavy machinery; it can change all of its inputs. In the long run the firm has no fixed inputs; we say all inputs are *variable*.

Note that the short run and the long run do not correspond to any particular length of time. Some industries may change their fixed inputs over weeks or months while others may do so over many years.

We will make extensive use of the distinction between the short run and the long run in Chapter 7. For now you should note that the ideas developed in this section are concerned *only with the short run.*

The meaning of marginal product

We will now examine the relationship between inputs and output in the short run. Since we are studying the short run, we know the firm has both fixed and variable inputs. For simplicity, let's consider a farm that produces potatoes, where the size of the farm and the agricultural machinery and tools are fixed. The variable inputs are labour and other agricultural inputs (seeds, fertiliser and so on). The only way the farmer can increase the quantity of output in the short run is by increasing the quantity of the variable inputs it uses. We can now distinguish between:

- **total product**, which is the total quantity of output (potatoes) produced by the firm
- **marginal product**, which is the extra or additional output (potatoes) produced by one additional unit of a variable input, which we will assume to be labour; it tells us by how much output increases as labour increases by one worker.

Table 2.4 shows how the total product and marginal product of each worker change as the number of workers increases. We can see that the total product increases continuously. The marginal or extra product added by each worker is simply the addition to total product by each worker. So the first worker adds 20 kilos of potatoes to total product, the second worker adds 30 kilos (= 50 – 20), the third worker 40 kilos (= 90 – 50) and so on.

Number of workers (the variable input)	Total product (kilos of potatoes)	Marginal product (kilos of potatoes)
0	0	-
1	20	20
2	50	30
3	90	40
4	120	30
5	140	20
6	150	0

Table 2.4: Total product and marginal product

Diminishing marginal returns

What is interesting to note in Table 2.4 is that the marginal product of our variable input of labour at first increases, it reaches a maximum with the third worker, and then begins to fall. This pattern of increasing and then falling marginal product is so universally valid it has the status of a law: it is known as the *law of diminishing marginal returns* or for short, the *law of diminishing returns*.

Let's examine the reasoning behind this pattern. When there are zero workers on the land, there is no output at all; it is equal to zero. When one worker is hired, there will be some output and so total product is 20 kilos of potatoes. Marginal product is also equal to 20 kilos. But one worker alone on the farm must do all the ploughing, planting, harvesting, and so on, and so output is quite low. When a second worker is hired, the two workers share the work, and total product increases to fifty kilos, indicating that the output produced by the two together is more than double the output of the first working alone. The additional (or marginal) product due to the second worker (30 kilos) is greater than that of the first (20 kilos). This process is repeated with the addition of the third, and marginal product increases.

With three workers, marginal product is the greatest it can be; when the fourth worker is added, marginal product begins to fall, and falls continuously thereafter. This is the point at which diminishing returns begin. Why does this happen? It happens because of overcrowding: each additional worker has less and less land to work with, and so produces less and less output. Eventually, the conditions on the farm become so crowded that the sixth worker adds zero extra output.

More generally, marginal product will begin to fall at some point not just on a farm with a fixed piece of land, but whenever more and more units of a variable input are added to a fixed input (provided the technology of production is unchanging). For example, in the case of a factory where more and more workers are hired, each extra worker will have fewer and fewer machines and equipment to work with, and so will add less and less output.

Imagine what would happen if diminishing returns did not exist. Using our farm example, it would be possible for food production to increase indefinitely just by continuously adding variable inputs to a fixed piece of land – a clear impossibility!

According to the **law of diminishing marginal returns** as more and more units of a variable input (such as labour) are added to one or more fixed inputs (such as land), the marginal product of the variable input at first increases, but there comes a point when it begins to decrease. This relationship presupposes that the fixed input(s) remain fixed, and that the technology of production is also fixed.

Increasing marginal costs and the firm's supply curve

We will now discover the relationship between diminishing marginal returns, costs of production, and the firm's supply curve.

The meaning of marginal costs

In economics we make frequent use of the concept of **total cost**, which refers very simply to all costs of production incurred by a firm. We can now consider the meaning of a new concept, *marginal cost*.

REAL WORLD FOCUS 2.1

David Ricardo and the end of agricultural output growth

David Ricardo, a famous English economist of the 19th century, believed that agricultural output would eventually stop growing, because as more and more labour and capital inputs were added to land that was fixed in quantity, the additional output of labour and capital would become smaller and smaller until it would no longer be possible for total output to increase further.

Figure 2.9: Chitwan, Nepal. People working in rice field at sunrise

Applying your skills

1. Explain the concept that describes the process Ricardo was referring to.
2. To what extent do you think Ricardo's fears were justified?
3. Suggest reasons for the growth of agricultural output in the real world in spite of a fixed quantity of land.

Marginal cost *is the extra or additional cost of producing one more unit of output*. It tells us by how much total costs increase if there is an increase in output by one unit.

To understand what this means, we can examine Table 2.5. We can see that as the number of units of total product increases, total cost increases, which is what we would expect since more output always involves more cost. Marginal cost shows the addition to total cost arising from the production of one more unit of output. So production of the first unit adds $12 to total cost, the second unit adds $8 to total cost, the third unit $6, and so on.

Total product (number of units)	Total cost ($)	Marginal cost ($)
1	12	12
2	20	8
3	26	6
4	34	8
5	46	12
6	62	16

Table 2.5: Total cost and marginal cost

How marginal costs are related to diminishing marginal returns

You may have noted that marginal cost at first decreases and then increases. In fact this is the opposite of what happens with diminishing marginal returns. This is not a coincidence, in fact the pattern followed by marginal cost is determined entirely by diminishing marginal returns. This can be seen in Figure 2.10.

When marginal product increases, marginal cost decreases; when marginal product is maximum, marginal cost is minimum; and when marginal product falls, marginal cost increases.

Let's examine this relationship more carefully. Remember that at low levels of output, the marginal product of labour, or the output of each extra worker increases. So each worker produces more and more output. Since workers add to costs, and each worker produces more and more output, the cost of producing each additional unit of output falls.

On the other hand, when marginal product or the additional output of each worker becomes less and less, the cost of each extra unit produced (marginal cost) must be increasing.

Similarly, when the additional output produced by an extra worker is the most it can be, then the extra labour

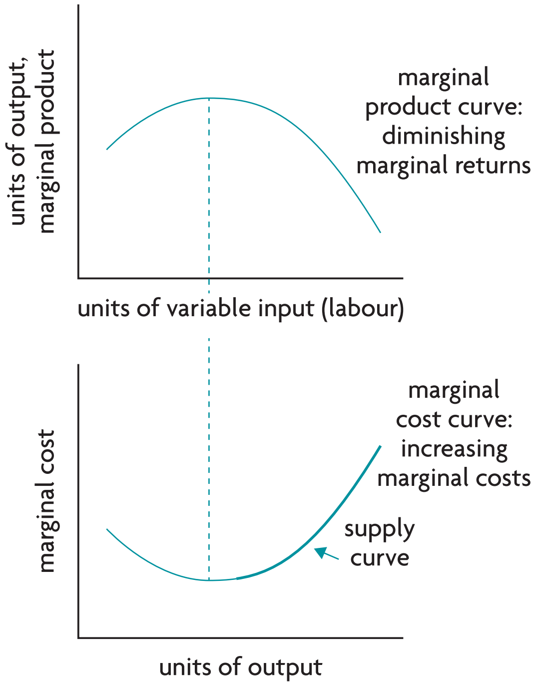


Figure 2.10: Marginal product, marginal cost and the firm's supply curve

cost of producing an additional unit of output is the least it can be.

Marginal cost and the firm's supply curve

As we discussed earlier, the firm's supply curve shows the quantities of a good the firm is willing and able to produce and sell at various prices, *ceteris paribus*. We now know, however, that the firm has costs of production. These costs must be covered through the firm's earnings from selling its output. The firm's earnings per unit of output sold are determined by the price of each unit of the good.

It follows then that the firm will be willing and able to supply some quantity as long as the price is enough to cover its costs. Therefore we can think of the supply curve as showing the price that the firm is willing to accept to produce one more unit of the good. In fact, when marginal cost is increasing, the firm can only produce more output if the price of the good increases to cover the extra cost of each extra unit produced. As a result, a portion of the upward sloping part of the marginal cost curve in Figure 2.10 is the firm's supply curve. This is represented in the bold face portion of the marginal cost curve. The supply curve begins at the point on the marginal cost curve where the firm is making enough revenue so that it is better off producing than shutting down.[1]

> The firm's supply curve is a portion of its marginal cost curve that shows the price-quantity combinations where the extra cost of producing one more unit of output (the marginal cost) is equal to the price of that unit.

TEST YOUR UNDERSTANDING 2.5

1. Explain the law of diminishing marginal returns and outline why this law holds only in the short run.
2. Explain why marginal cost falls as marginal product increases, and why it increases as marginal product falls.
3. Analyse the relationship between a firm's marginal costs and its supply curve.

2.4 Competitive market equilibrium: demand and supply

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- draw a diagram to illustrate how demand and supply interact to determine market equilibrium (AO4)
- analyse demand curve and supply curve shifts that give rise to a new market equilibrium, using the concepts of excess demand (shortage) and excess supply (surplus) (AO2)
- draw diagrams to illustrate how shifts in demand and supply give rise to a new equilibrium (AO4)

[1] This occurs when the firm is covering all of its variable costs. An explanation of this topic is beyond the scope of the IB syllabus.

The market demand and market supply for chocolate bars that we considered separately earlier in this chapter show the quantities consumers and firms are *willing and able* to buy and sell at each price. We will now put market demand and market supply together to find out how these interact to determine what happens in the market for chocolate bars.

Market equilibrium

Excess demand (shortages) and excess supply (surpluses)

Figure 2.11 presents the same market demand and supply curves that appeared in Figures 2.2(c) and 2.6(c). The same information appears as a demand schedule and a supply schedule in Table 2.6.

In both Table 2.6 and Figure 2.11 we see that when the price of chocolate bars is $3, quantity demanded is exactly equal to quantity supplied, at 8000 chocolate bars. Note that there is only one price where this can occur. At a higher price, say $4, quantity supplied (10 000 bars) is greater than quantity demanded (6000 bars). There is **excess supply**, or a **surplus** of 4000 bars (10 000 – 6000). At the even higher price of $5, there is a larger excess supply (surplus) of 8000 bars.

Suppose the price in this market is initially $5. At this price, chocolate producers would be willing and able to produce 12 000 bars, but consumers would only be willing and able to buy 4000 bars. What will happen? With unsold output of 8000 bars, producers will lower their price to encourage consumers to buy more chocolate. As the price falls, quantity supplied becomes smaller and quantity demanded becomes bigger. As long as there is a surplus, there will be a downward pressure on the price. The price will keep falling until it reaches the point where quantity demanded is equal to quantity supplied, and the surplus is eliminated.

At a lower price than $3, say $2, quantity demanded (10 000 bars) is larger than quantity supplied (6000 bars). There is now **excess demand** or a **shortage** of 4000 bars (10 000 – 6000). If price were even lower, at $1, the shortage would be larger, at 8000 bars. At a price of $1, producers would be willing and able to supply only 4000 bars, whereas consumers would be willing and able to buy 12 000 bars. Producers will notice that the chocolate bars are quickly sold out, and so begin to raise the price. As they do so, quantity demanded begins to fall and quantity supplied begins to rise. The shortage in the chocolate market exerts an upward pressure on price. The price will keep increasing until the shortage is eliminated; this will happen when quantity supplied is exactly equal to quantity demanded.

Price of chocolate bars ($)	Quantity of chocolate bars demanded (per week)	Quantity of chocolate bars supplied (per week)
5	4000	12 000
4	6000	10 000
3	8000	8000
2	10 000	6000
1	12 000	4000

Table 2.6: Market demand and supply schedules for chocolate bars

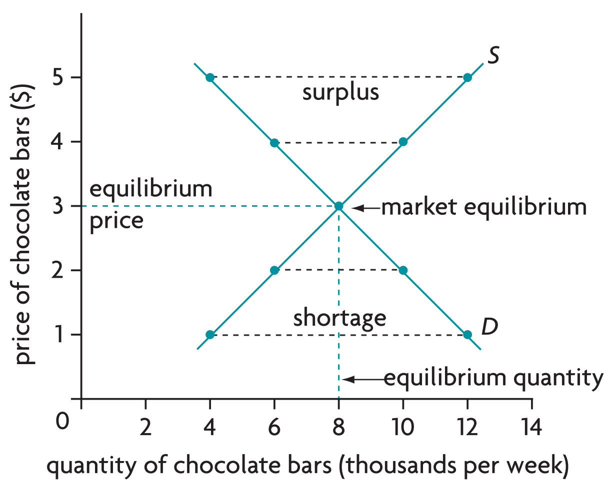


Figure 2.11: Market equilibrium

The existence of *excess demand* (a *shortage*) or *excess supply* (a *surplus*) in a free market will cause the price to change so that the quantity demanded will be made equal to quantity supplied. In the event of excess demand, price will rise; in the event of excess supply, price will fall.

Market equilibrium

Equilibrium is defined as a state of balance between different forces, such that there is no tendency to change. This is an important concept in economics that we will encounter repeatedly. When quantity demanded is equal to quantity supplied, there is **market equilibrium**; the forces of supply and demand are in balance, and there is no tendency for the price to change. Market equilibrium is determined at the point where the demand curve intersects the supply curve. The price in market equilibrium is the **equilibrium price**, and the quantity is the **equilibrium quantity**. At the equilibrium price, the quantity consumers are willing and able to buy is exactly equal to the quantity firms are willing and able to sell. This price is known as the *market price*. In the market for chocolate bars in Figure 2.11, the equilibrium price is $3 per chocolate bar, and the equilibrium quantity is 8000 bars. At any price other than the equilibrium price, there is *market disequilibrium*. In a free competitive market, a market disequilibrium cannot last, as demand and supply force the price to change until it reaches its equilibrium level.

> **competitive market equilibrium**, quantity demanded equals quantity supplied, and there is no tendency for the price to change. In a market disequilibrium, there is excess demand (shortage) or excess supply (surplus), and the forces of demand and supply cause the price to change until the market reaches equilibrium.

Changes in market equilibrium

Once a price reaches its equilibrium level, consumers and firms are satisfied and will not engage in any action to make it change. However, if there is a change in any of the non-price determinants of demand or supply, a shift in the curves results, and the market will adjust to a new equilibrium.

Changes in demand (demand curve shifts)

In Figure 2.12(a), D_1 intersects S at point a, resulting in equilibrium price and quantity P_1 and Q_1. Consider a change in a determinant of demand that causes the demand curve to shift to the right from D_1 to D_2 (for example, an increase in consumer income in the case of a normal good). Given D_2, at the initial price, P_1, there is a movement to point b, which results in excess demand equal to the horizontal distance between points a and b. Point b represents a disequilibrium, where quantity demanded is larger than quantity supplied, thus exerting an upward pressure on price. The price therefore begins to increase, causing a movement up D_2 to point c, where excess demand is eliminated and a new equilibrium is reached. At c, there is a higher equilibrium price, P_2, and greater equilibrium quantity, Q_2, given by the intersection of D_2 with S.

A decrease in demand, shown in Figure 2.12(b), leads to a leftward shift in the demand curve from D_1 to D_3 (for example, due to a decrease in the number of consumers). Given D_3, at price P_1, there is a move from the initial equilibrium (point a) to point b, where quantity demanded is less than quantity supplied, and therefore a disequilibrium where there is excess supply equal to the horizontal difference between a and b. This exerts a downward pressure on price, which

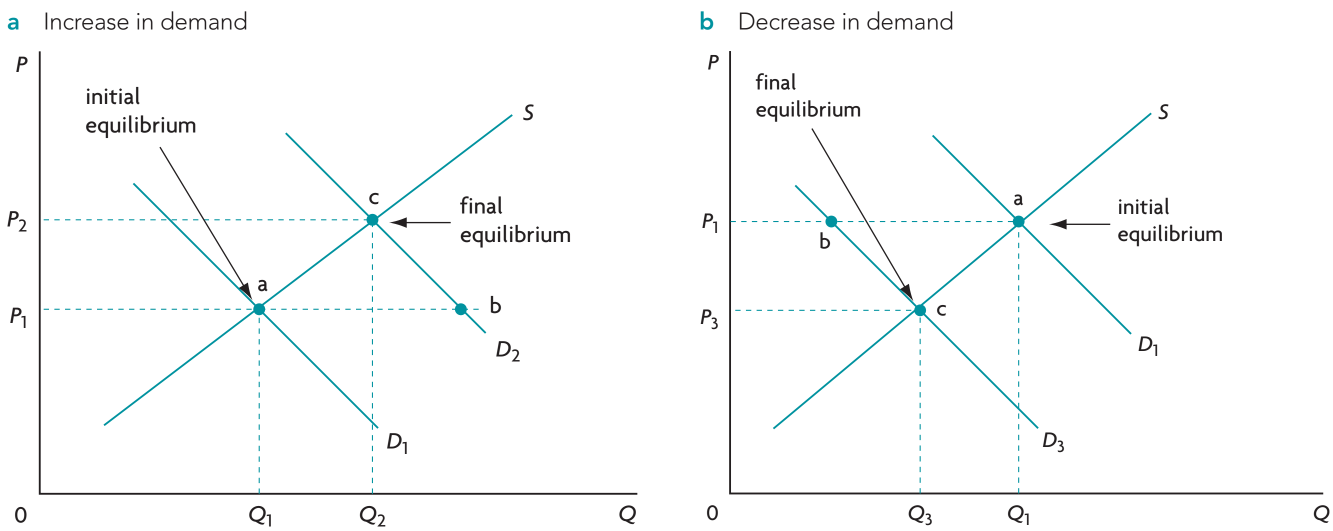


Figure 2.12: Changes in demand and the new equilibrium price and quantity

falls, causing a movement down D_3 to point c, where excess supply is eliminated, and a new equilibrium is reached. At c, there is a lower equilibrium price, P_3, and a lower equilibrium quantity, Q_3, given by the intersection of D_3 with S.

Changes in supply (supply curve shifts)

We now consider supply curve shifts that can arise from changes in the determinants of supply. In Figure 2.13(a), the initial equilibrium is at point a where D intersects S_1, and where equilibrium price and quantity are P_1 and Q_1. An increase in supply (say, due to an improvement in technology) shifts the supply curve to S_2. With S_2 and initial price P_1, there is a move from point a to b, where there is disequilibrium due to excess supply (by the amount equal to the horizontal distance between a and b). Therefore, price begins to fall, and there results a movement down S_2 to point c where a new equilibrium is reached. At c, excess supply has been eliminated, and there is a lower equilibrium price, P_2, but a higher equilibrium quantity, Q_2.

A decrease in supply is shown in Figure 2.13(b) (say, due to a fall in the number of firms). With the new supply curve S_3, at the initial price P_1, there has been a move from initial equilibrium a to disequilibrium point b, where there is excess demand (equal to the distance between a and b). This causes an upward pressure on price, which begins to increase, causing a move up S_3 until a final equilibrium is reached at point c, where the excess demand has been eliminated, and there is a higher equilibrium price P_3 and lower quantity Q_3.

Free and economic goods revisited

You may remember from Chapter 1, Section 1.1, that free goods are goods that are not subject to the condition of scarcity and have a zero opportunity cost, while economic goods are subject to scarcity and have an opportunity cost greater than zero. We are now in a better position to understand the difference between the two by use of demand and supply analysis. Figures 2.14(a) and (b) show a free and economic good respectively. A *free good* is a good for which the quantity supplied is greater than the quantity demanded when the price is zero. Supply is so large relative to demand that there is excess quantity supplied even at a zero price. An *economic good* is a good for which quantity supplied is smaller than quantity demanded when the price is zero. A free good can change into an economic good as a result of a leftward shift in the supply curve (reduced supply) or a rightward shift in the demand curve (increased demand). When demand and supply intersect at a price greater than zero, the good has become an economic good.

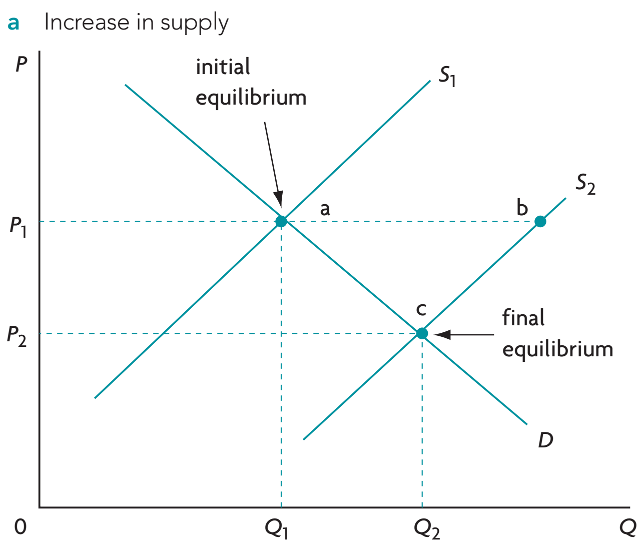
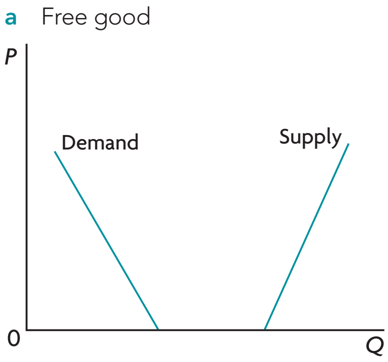


Figure 2.13: Changes in supply and the new equilibrium price and quantity



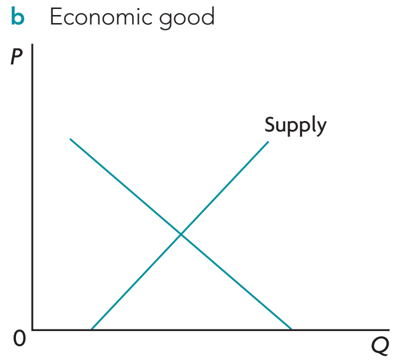


Figure 2.14: Free and economic goods

REAL WORLD FOCUS 2.2

Part I: Avocados and demand for healthy fats

For thousands of years the avocado was held to be a sacred fruit in Central America.Yet this was not so in much of the rest of the world. In 1982, it was reported that an oversupply of avocados in the United States led to a drop in price so large that producers considered selling it as food for dogs (though avocados cause an upset stomach in dogs). An important reason why consumers avoided avocados was their high fat content.

In more recent years, consumers' regard for the avocado has changed dramatically, thanks to the distinction that science has made between the 'good' and 'bad' fat content of foods. Avocados are plentiful in 'good' or healthy (monounsaturated) fats, and in addition contain a host of healthful minerals and vitamins. Consumers around the world are now eating far more avocados than ever before, and their price is rising.

Source: *https://riverheadlocal.com*

Figure 2.15: Green Hass avocado fruit

Part II: Potato chips (crisps) and the supply of potatoes

The summer of 2018 was the hottest in the UK since 1976. As a result there was a significant drop in quantities produced of many crops. By the fall UK farmers noted that potato crops were reduced by one-third. This gave rise to an increase in the price of potato chips (crisps).

Source: *www.independent.co.uk*

Applying your skills

Using diagrams in each case, analyse and explain

1 the effect on avocado prices of consumers' perceptions that avocados are unhealthy due to their high fat content

2 the effect on avocado prices of the information that avocados are very healthy

3 the effect on the price of potato chips (crisps) of extreme weather conditions in the UK.

TEST YOUR UNDERSTANDING 2.6

1 In Figure 2.11, state whether there is excess supply (a surplus) or excess demand (a shortage), and how large this is if price per chocolate bar is:

a $5,

b $4,

c $3,

d $2, and

e $1.

2 Use a demand and supply diagram to:

a show equilibrium price and quantity,

b show possible disequilibrium prices and quantities,

c relate disequilibrium prices to excess demand (shortages) and excess supply (surpluses),

d explain the meaning of 'market equilibrium', and

e explain the roles of demand and supply in achieving market equilibrium.

3 Use supply and demand diagrams to illustrate the following events.

a Freezing weather destroys the orange crop and the price of oranges rises.

b The mass media report on the fat content of cheese and the price of cheese falls.

c A new technology of production for computers is developed and the price of computers falls.

d Milk, an input for ice cream production, becomes more expensive and the price of ice cream increases.

e The mass media report on the health benefits of olive oil and the price of olive oil increases.

4 Assuming a competitive market, use demand and supply diagrams to show in each of the following cases how the change in demand or supply for product A creates a disequilibrium consisting of excess demand or excess supply, and how the change in price eliminates the disequilibrium.

a Consumer income increases (A is a normal good).

b Consumer income falls (A is an inferior good).

c There is an increase in labour costs.

d The price of substitute good B falls.

e The number of firms in the industry producing product A increases.

f A successful advertising campaign emphasises the health benefits of product A.

2.5 The role of the price mechanism and market efficiency

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- analyse the price mechanism with respect to its functions including (AO2)
  - resource allocation (signalling and incentive functions)
  - rationing

CONTINUED

- explain the meaning of consumer and producer surplus, as well as social or community surplus (AO2)
- draw a diagram showing maximum social surplus at competitive market equilibrium (AO4)
- explain the conditions for achievement of allocative efficiency at competitive market equilibrium (AO2)
  - social surplus is maximum
  - marginal benefits are equal to marginal costs
- calculate consumer and producer surplus from a diagram (AO4)

Prices determined by the forces of supply and demand in competitive markets are known as the **price mechanism**. The price mechanism serves some important functions that we will now explore.

Functions of the price mechanism

Prices as signals and incentives and the allocation of resources

In Chapter 1 we saw that the condition of scarcity forces societies to make choices. As the production possibilities model illustrated (Figure 1.1), assuming the economy is producing on its production possibilities curve (*PPC*), it must decide on what particular point on the *PPC* it wishes to produce. This involves a choice about *what/how much to produce*, which is a decision on one aspect of resource allocation. It also involves a choice about which of its available resources and in what quantities, it will allocate to produce the combination of goods and services it has chosen. This is a choice on the *how to produce* question of resource allocation.

This brings us to an important question. How does a society make a choice about where to be on its *PPC*? Who decides, and how is this decision carried out? In a market economy, it is simply prices in free markets, resulting from the interactions of demanders and suppliers, which make the decisions and carry them out. We have learned that when markets operate under competitive conditions, market demand and market supply, composed of numerous individual demanders and suppliers, determine equilibrium prices and quantities for goods (and services and resources). At these equilibrium positions, the buying and selling choices of all buyers and sellers are satisfied and are in balance. This market mechanism, working through prices, is known as the *invisible hand of the market*, a phrase first used by Adam Smith, whose contributions to economics were explained in Chapter 1. The invisible hand succeeds in co-ordinating the buying and selling decisions of thousands or millions of decision-makers in an economy without any central authority. The *what/how much to produce* question of resource allocation is answered because firms produce only those goods consumers are willing and able to buy, while consumers buy only those goods producers are willing and able to supply; and the *how to produce* question of resource allocation is answered because firms use those resources and technologies in their production process that they are willing and able to pay for.

How do prices and markets achieve the task of resource allocation?

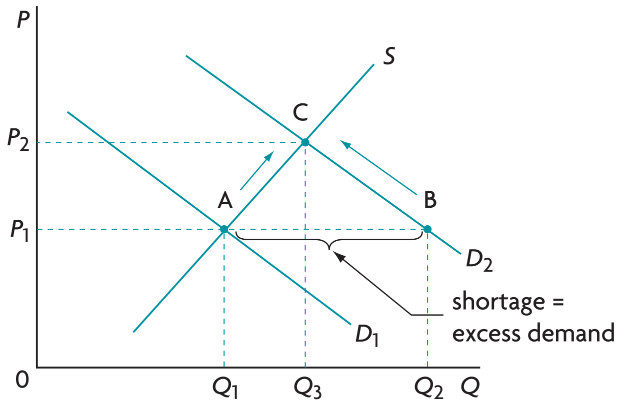
The key to the market's ability to allocate resources can be found in the **signalling** and **incentive** functions of prices in resource allocation. As signals, prices communicate information to decision-makers. As incentives, prices motivate decision-makers to respond to the information.

We will examine the signalling and incentive functions of prices by use of the following two examples.

An example from a product market

Suppose consumers decide they would like to eat more strawberries because of their health benefits (a change in tastes); demand increases and the demand curve shifts to the right from D_1 to D_2 in Figure 2.16(a). At the initial price, P_1, this results in a shortage equal to the difference between Q_2 and Q_1: the quantity demanded Q_2, due to the increase in demand to D_2, is larger than quantity supplied, Q_1. The price of strawberries therefore begins to rise and will continue to rise until the shortage has disappeared. This happens at price P_2 and quantity Q_3, given by the point of intersection of the supply curve with the new demand curve, D_2.

a Adjustment of price to increased demand



b Adjustment of the price of labour to increased labour supply

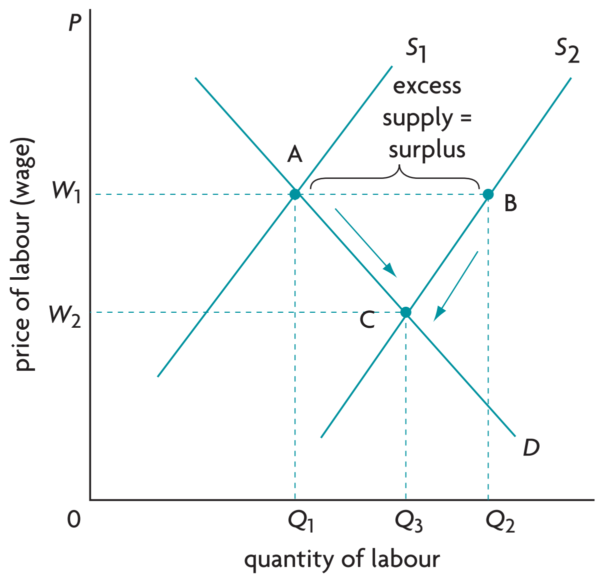


Figure 2.16: Price as a signal and incentive

What has happened? The new, higher price signalled or conveyed information to producers that a shortage in the strawberry market had emerged. The increase in price is also an incentive for producers to increase the quantity of strawberries supplied; at the higher price, strawberry production is more profitable, so producers move along the supply curve from point A to point C, increasing quantity supplied from Q_1 to Q_3. But the new, higher price is a signal and incentive for consumers: it signals that strawberries are now more expensive, and is an incentive for them to buy fewer strawberries. They therefore move along the new demand curve from B to C, buying fewer strawberries than at the original price P_1 (Q_3 is smaller than Q_2). The increase in the price of strawberries resulted in a *reallocation of resources*. More resources are now allocated to strawberry production. (This affects the answer to the *what to produce* question of resource allocation.)

An example from a resource market

The second example is from the labour market (a resource market). The vertical axis in Figure 2.16(b) measures the price of labour (the wage) and the horizontal axis the quantity of labour. Firms are interested in buying labour services, and their demand for labour is given by *D*. Owners of labour services (workers) supply their labour in the labour market, and the initial supply of labour is shown by S_1.[2]

Assume that because of immigration (foreign workers enter the country), the supply of labour increases, so the labour supply curve shifts to S_2. At the old wage, W_1, there is a surplus of labour shown by the difference between Q_2 and Q_1 of labour. The surplus causes the wage to start falling, and this falls until the surplus has disappeared. The new equilibrium wage is W_2, and the equilibrium quantity of labour Q_3, given by the intersection of *D* with S_2.

The falling wage has acted as a signal and an incentive. It signalled to firms that there was a surplus in the labour market, and it provided them with an incentive to hire more labour; therefore, they move along the labour demand curve from point A to point C (Q_3 is larger than Q_1). The lower wage is also a signal to workers, providing them with the incentive to move along the new supply curve, S_2, from B to C, where they offer less of their services at the lower wage (Q_3 is less than Q_2). With firms and workers responding to price signals and incentives, there occurred a reallocation of labour resources with firms now producing output with a larger quantity of labour. (This affects the answer to the *how to produce* question of resource allocation.)

[2] The demand curve has the usual downward-sloping shape, because as the wage falls, firms are prompted to hire more labour and so the quantity of labour demanded increases. The supply curve has the usual upward-sloping shape because the higher the wage, the more willing workers will be to supply their labour in the market.

Prices and rationing

Rationing is a method of apportioning or parceling out goods and services among consumers or households. The market mechanism uses *price rationing* for this purpose, which involves the use of prices freely determined in markets. This simply means that whether or not a consumer will get a good is determined by the price of that good. All those who are willing and able to pay the price of a good or service will get it, and all those who are not willing or able to pay the price will not get it. Price, and price alone is the factor that decides. This happens because at the market-determined equilibrium price and quantity, the buying and selling choices of all buyers and sellers in the market for a good are satisfied and are in balance.

Price rationing is simply the market mechanism at work, known as Adam Smith's *invisible hand of the market* (see Chapter 1), which coordinates the countless buying and selling decisions of decision-making units in an economy without any central authority.

When the price mechanism is not working properly, such as in planned economies (see Chapter 1) or if there are price controls (see Chapter 4), then *non-price rationing* systems come into play in order to apportion goods and services among their various users.

The most common non-price rationing method is the waiting line or queue, where consumers get a good on a first come, first served basis. Non-price rationing will be discussed in Chapter 4.

TEST YOUR UNDERSTANDING 2.7

1 a Explain the role of prices in resource allocation.

b Describe how this relates to Adam Smith's *invisible hand*.

2 Consider the market for coffee, and suppose that the demand for coffee falls (because of a fall in the price of tea, a substitute good), leading to a new equilibrium price and quantity of coffee. Using diagrams, explain the role of price as a signal and as an incentive for consumers and for firms in reallocating resources.

3 Consider the labour market, and suppose the supply of labour falls (due to large-scale departure of workers to another country), resulting in a new equilibrium price and quantity of labour. Using diagrams, explain the role of price as a signal and as an incentive for workers (the suppliers of labour) and firms (the demanders of labour) in reallocating resources.

4 Distinguish between price rationing and non-price rationing.

Allocative efficiency in competitive markets

Efficiency, as we saw in Chapter 1, is one of the key concepts running through this course, and very broadly means making the best possible use of resources. We will now study how competitive markets achieve a special type of efficiency called allocative efficiency.

Allocative efficiency refers to producing the quantity of goods mostly wanted by society. Allocative efficiency is achieved when the economy allocates its resources so that the society gets the most benefits from consumption. Since allocative efficiency refers to producing what consumers mostly want, it answers the *what/how much to produce* question in the best possible way.[3]

Reinterpreting demand and supply: marginal benefits and marginal costs

Before we consider allocative efficiency at competitive market equilibrium, we will pause a moment to explore a different interpretation of demand and supply.

The demand curve as a marginal benefit curve

Consumers buy goods and services because these provide them with some benefit, or satisfaction. The greater the quantity of a good consumed, the greater the satisfaction. However, the extra benefit provided by each additional unit increases by smaller and smaller amounts. Imagine you buy a soft drink, which provides you a certain amount of benefit. You are still thirsty, so you buy a second soft drink. Whereas you will enjoy this you will most likely enjoy it less than

[3] At the same time it also answers the *how to produce* question in the best possible way. An explanation of this is beyond the scope of the IB syllabus.

the first; the second soft drink provides you with less benefit than the first. If you buy a third, you will get even less benefit than from the second, and so on with each additional soft drink. The extra benefit that you get from each additional unit of something you buy is called the **marginal benefit** (marginal means extra or additional).

Now remember that the demand curve, as you learned earlier in this chapter, shows the various quantities of a good that a consumer is willing and able to buy at different possible prices, *ceteris paribus*. But it can also be thought of as showing the amount of money that the consumer is willing to pay to get one more unit of the good. Since the extra or marginal benefit that the consumer gets from buying one more unit of the good becomes smaller and smaller as the quantity increases, this means that the price the consumer is willing to pay also gets smaller and smaller as the number of units bought increases.

Since marginal benefit decreases as the quantity of a good consumed increases, consumers will be willing to buy an extra unit of the good only if its price falls. The demand curve can therefore also be called a *marginal benefit (MB) curve*

(Students studying this course at HL will notice that the concept of marginal benefits is very similar to the concept of marginal utility examined earlier. The difference between the two is actually very subtle. Marginal utility as we have seen is something that cannot be measured. Marginal benefit, on the other hand, is interpreted to indicate the consumer's *willingness to pay* for the last or marginal unit bought, which is simply price, and is therefore measurable. The differences between the two concepts are beyond the scope of the IB syllabus.)

The supply curve as a marginal cost curve

In order to be able to produce goods and services, firms have *costs*, which involve payments that must be made in order to acquire factors of production. For example, a farmer must buy seeds and fertilisers, pay workers to work on the land, and so on. The extra cost of producing one more unit of output is called the *marginal cost*. Marginal cost typically increases as the units of output produced increase. (The reasons for this were studied at HL earlier in this chapter in Section 2.3.) This means that the farmer can only produce more output if the price of the good increases to cover the extra cost of each extra unit produced.

The firm's supply curve as you may remember, shows the various quantities of a good that a firm is willing and able to produce and supply at various possible prices, *ceteris paribus*. But in view of the firm's costs of production, the supply curve also shows the price that the firm is willing to accept in order to produce one more unit of the good.

Since marginal cost increases as the quantity of a good produced increases, producers will be willing to produce and sell an extra unit of the good only of its price increases. The supply curve can therefore also be called a *marginal cost (MC) curve*.

Allocative efficiency: *MB* = *MC* at competitive market equilibrium

As we know, market equilibrium occurs at the point of intersection of the demand and supply curves, but depending on how we interpret the demand and supply curves, market equilibrium can be thought of differently. If we interpret the demand curve as a marginal benefit (MB) curve, and the supply curve as a marginal cost (MC) curve, then market equilibrium occurs where $MB = MC$. The equality of MB with MC tells us that the extra benefit to society of getting one more unit of the good is equal to the extra cost to society of producing one more unit of the good. When this happens, society's resources are being used to produce the 'right' quantity of the good; in other words, society has allocated the 'right' amount of resources to the production of the good, and is producing the quantity of the good that is mostly wanted by society. *This is none other than allocative efficiency* and is shown in Figure 2.17.

To understand this, consider that if $MB > MC$, then society would be placing a greater value on the last unit of the good produced than it costs to produce it, and so more of it should be produced. If $MC > MB$, then it would be costing society more to produce the last unit of the good produced than the value society puts on it, and so less should be produced. If $MC = MB$, then just the 'right' quantity of the good is being produced.

Putting the above points together, we can conclude that at the point of competitive market equilibrium, where $MB = MC$, the economy achieves allocative efficiency. For allocative efficiency to be achieved for an entire economy, the condition $MB = MC$, must hold in all markets.

Introducing consumer and producer surplus

There is another way we can understand how allocative efficiency is achieved by the competitive market economy; this involves the concepts of consumer surplus and producer surplus.

Consumer surplus

Consumer surplus is defined as the highest price consumers are willing to pay for a good minus the price actually paid. The highest price they are willing to pay is given by the demand curve. The price actually paid is determined at the market equilibrium by supply and demand. Consumer surplus is shown in Figure 2.17 as the shaded area between the demand curve, and the equilibrium price P_e, up to quantity Q_e.

Consumer surplus is the area under the demand curve and below the price paid by the consumer, up to the quantity purchased.

Consumer surplus indicates that whereas many consumers were willing to pay a higher price to get the good, they actually received it for less. For example, many consumers were willing to pay price P_2 to get quantity Q_a, yet they got Q_a by paying only the lower price P_e. The difference between P_2 and P_e is consumer surplus for quantity Q_a. Similarly, many consumers were willing to pay price P_3 in order to get quantity Q_b, yet they got it by paying only P_e. Again, the difference $P_3 - P_e$ is consumer surplus for quantity Q_b. The same principle applies to all possible prices between the highest price P_1 and the equilibrium price P_e. Therefore, all the consumers who were willing to pay a higher price than P_e to get the good received some benefit over and above what they actually paid for the good. This extra benefit is called consumer surplus.

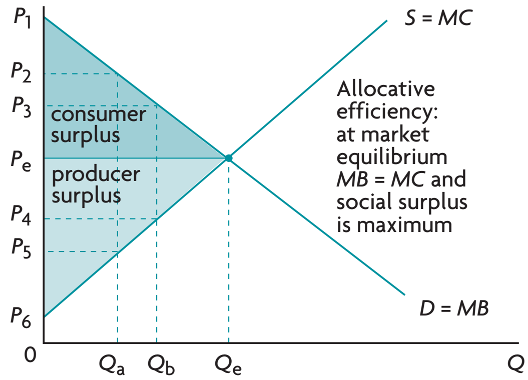


Figure 2.17: Consumer and producer surplus in a competitive market

Producer surplus

Producer surplus is defined as the price received by firms for selling their good minus the lowest price that they are willing to accept to produce the good. The price received is given by the market equilibrium price. The lowest price they are willing to accept is shown by the supply curve. Producer surplus appears in Figure 2.17 as the shaded area between the supply curve, and the equilibrium price P_e, up to quantity Q_e.

Producer surplus is shown as the area above the firms' supply curve and below the price received by the firm, up to the quantity produced.

As we can see in Figure 2.17, firms that were willing to produce quantity Q_a for price P_5 actually received price P_e. The difference $P_e - P_5$ is producer surplus for quantity Q_a. Similarly, the producer surplus for quantity Q_b is given by the price P_e actually received minus P_4 that the firms were willing to accept for producing Q_b. The same principle applies to all possible prices between the lowest price P_6 and the equilibrium price P_e. Therefore, total producer surplus is shown by the shaded area between the equilibrium price P_e and the supply curve, up to the quantity produced and sold.

Competitive market equilibrium: maximum social surplus and allocative efficiency

At the point of competitive market equilibrium, the sum of consumer and producer surplus is maximum, or the greatest it can be. To see why, consider what would happen if any quantity less than Q_e were produced in Figure 2.17. If, say, Q_b is produced, the sum of consumer plus producer surplus would be smaller, as this sum would be equal to the shaded area between the demand and supply curves *only up to output* Q_b. It follows, then, that the sum of consumer plus producer surplus is maximised at the point of market equilibrium. The sum of consumer and producer surplus is known as **social surplus** (or **community surplus**).[4]

At the point of competitive market equilibrium, social surplus, defined as the sum of consumer plus producer surplus, is maximum.

Let's now examine the importance of maximum social surplus at the point of competitive market equilibrium. You will note that competitive market equilibrium is none other than the point where marginal benefit equals marginal cost, or $MB = MC$. Based on our discussion above, we know that when $MB = MC$ there is allocative efficiency. Therefore maximum social surplus is just another way of saying that there is allocative efficiency.

At the point of competitive market equilibrium shown in Figure 2.17, production of a good occurs where $MB = MC$, which is also *where social surplus, or the sum of consumer plus producer surplus is maximum*. This means that markets are achieving allocative efficiency, producing the quantity of goods mostly wanted by society. Society is making the best possible use of its scarce resources.

When the competitive market realises allocative efficiency, we say that 'social welfare' is maximised. What does this mean? The term **welfare** in economics refers to the amount of consumer and producer surplus. Welfare is clearly maximum when social surplus is maximum, or when $MB = MC$. We can therefore say that:

In competitive markets, when $MB = MC$, or when social surplus is maximum, social welfare is maximum.

Calculating consumer and producer surplus and social surplus

Figure 2.18(a) is similar to Figure 2.17 only in that it includes figures that allow us to calculate consumer and producer surplus. To see how this is done, note the following two terms.

P intercept = the point on the P axis that is met by a curve. In Figure 2.18(a), the demand curve has a P intercept = 6.5 and the supply curve has a P intercept = 1.

Q intercept = the point on the Q axis that is met by a curve. In Figure 2.18(a), we are not shown any Q intercept. However, in Figure 2.18(b) we see that the supply curve has a Q intercept = 2.

Let's begin with a calculation of consumer and producer surplus in Figure 2.18(a). Consumer surplus, as we know, is the area under the demand curve and below the price paid by consumers, up to the quantity purchased. It is simple to calculate consumer surplus if we think of it as half the area of the rectangle whose one side equals the *P* intercept of the demand curve minus the price paid by consumers, and whose other side equals the number of units purchased.

Consumer surplus =

$$\frac{\left(P \text{ intercept of D curve minus } P \text{ of consumers}\right) \times Q \text{ purchased}}{2}$$

$$\text{Therefore consumer surplus} = \frac{(6.5-3)\times 4.5}{2}$$

$$= \frac{3.5\times 4.5}{2} = \$7.88$$

Producer surplus is the area above the supply curve and below the price received by producers, up to the quantity produced. In Figure 2.18(a), it is half the area of the rectangle whose one side equals the price received by producers minus the *P* intercept of the supply curve, and whose other side equals the number of units sold.

Producer surplus =

$$\frac{\left(P \text{ of producers minus } P \text{ intercept of } S \text{ curve}\right) \times Q \text{ sold}}{2}$$

[4] Note that the term 'surplus' is used in two senses in this chapter. In one sense it refers to 'excess supply' (Section 2.4) and in the other it refers to benefits received by consumers and producers. You can avoid confusing the two concepts by noting that whenever the term is used on its own, it refers to excess supply.

Therefore:

$$\text{producer surplus} = \frac{(3-1)\times 4.5}{2} = \frac{(2\times 4.5)}{2} = \$4.50$$

Notice that both consumer surplus and producer surplus are given in terms of $, this is because they express a *monetary value* that has been gained by consumers and producers.

It follows then that social surplus, being the sum of consumer and producer surplus = $7.88 + $4.50 = $12.38.

However, it should be noted that the supply curve does not always have a P intercept. This can be seen in Figure 2.18(b), where the supply curve begins on the Q axis, having a Q intercept = 2. In cases like this, the formula given above for producer surplus is not appropriate. Instead, to find the area above the supply curve and under the price up to the quantity produced we need to use the formula for a *trapezium*. A trapezium is a geometrical shape where one pair of opposite sides is parallel but the other pair is not parallel. In Figure 2.18(b), let's call the lengths of the two parallel sides a and b, while the distance between these two sides is called c. The area of a trapezium then is:

$$\text{area of trapezium} = \frac{(a+b)\times c}{2}$$

Therefore considering Figure 2.18(b) we can calculate that:

$$\text{producer surplus} = \frac{(4+2)\times 3}{2} = \frac{6\times 3}{2} = \$9.00$$

Consumer surplus in Figure 2.18(b), using the formula given earlier is:

$$\text{consumer surplus} = \frac{(6.5-3)\times 4}{2} = \frac{3.5\times 4}{2} = \$7.00$$

Social surplus, being the sum of consumer and producer surplus = $7.00 + $9.00 = $16.00.

As we will discover in later chapters, consumer surplus is always given by the area of a triangle. Producer surplus and social surplus are usually given by the area of a triangle but sometimes they may be the area of a trapezium. Therefore, before you use a formula for a calculation, you must first identify whether you are calculating the area of a triangle or the area of a trapezium.

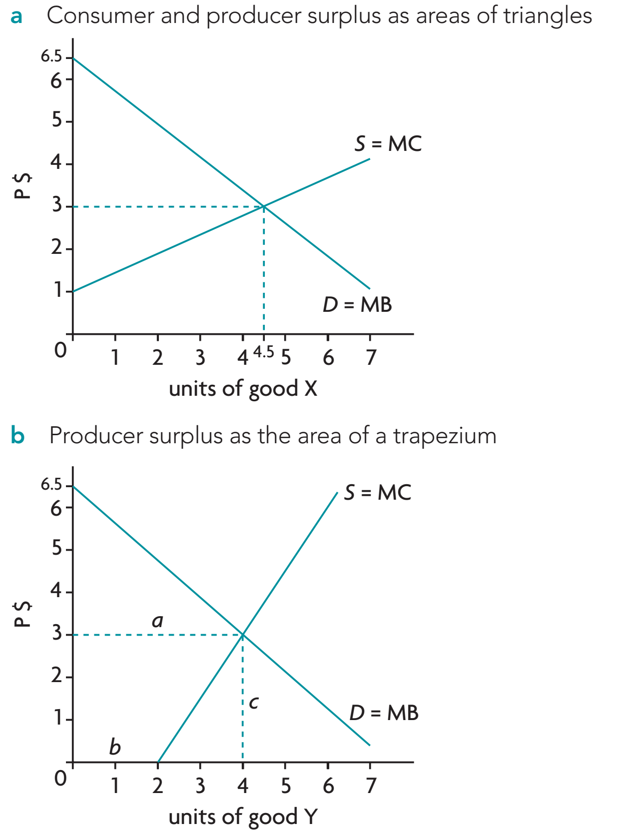


Figure 2.18: Calculating consumer and producer surplus and social surplus

Adam Smith's invisible hand

Adam Smith did not think of consumer and producer surplus, or marginal benefits and marginal costs, but he did come up with the idea that individuals acting in the best self-interest would through the workings of markets, produce an outcome for society that would be in society's best interests. We are now in a far better position to understand the *invisible hand* that was introduced in Chapter 1. It means that the market is able to coordinate the decisions of countless actions of individual economic decision-makers without any central authority, simply through the workings of demand and supply, while at the same time promoting efficiency which encourages the best allocation of scarce resources. This is perhaps the most important contribution of Adam Smith to economic thought.

Allocative efficiency of markets versus government intervention

We have seen that the competitive market succeeds in achieving allocative efficiency, thus ensuring the best possible use of scarce resources. This idea suggests that there should not be government intervention in markets, as these work very well on their own. However, there are two important issues that arise, calling into question the idea that governments should not intervene.

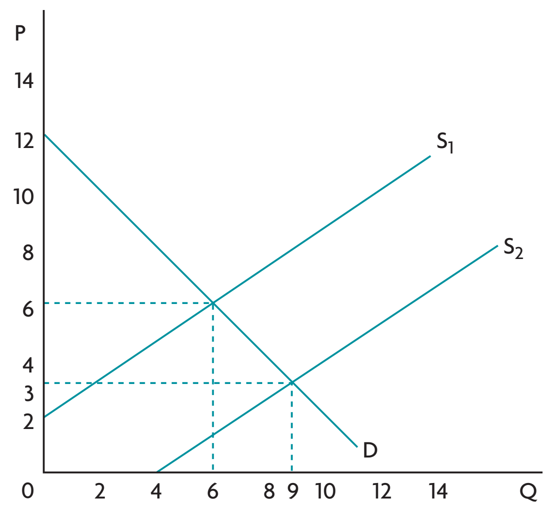
The first is that efficiency can only arise under a number of very strict and highly unrealistic conditions that are practically never met in the real world. In the real world the market fails to achieve allocative efficiency. Market failures are an important justification for government intervention. When markets fail to achieve allocative efficiency, this means that social surplus is reduced, resulting in what is known as **welfare loss** (also called *deadweight loss*). In other words, social welfare is no longer maximum on account of a portion of it being lost. Welfare loss will be discussed in later chapters.

The second is that the competitive market is unable to provide a satisfactory answer to the *for whom to produce* question, or output and income distribution, thus also inviting some government intervention. The topic of distribution and what can be done to improve outcomes will also be examined in Chapters 12 and 19, as well as in Chapter 6 at HL. It is also discussed in the Theory of knowledge feature 2.1.

These observations do not lessen the significance of the market's potential advantages; they only point out that in the real world, there is a need for government policies to counteract the failings of markets, thus allowing them to realise their potential advantages. There are important reasons why economists study the competitive market extensively. One is that government policies can be assessed with respect to their efficiency consequences (Chapter 4). Another is that it can form the basis for government policies that try to create conditions in the real world that allow actual economies to come closer to achieving economic efficiency (Chapters 5–7). A third is that it provides standards for economic efficiency against which actual outcomes, which are less than perfectly efficient, can be assessed (Chapters 4 and 7).

TEST YOUR UNDERSTANDING 2.8

1. Explain the meaning of allocative efficiency.
2. Explain the meaning of
   a. consumer surplus,
   b. producer surplus,
   c. social (or community) surplus, and
   d. maximum 'social welfare'.
3. a. Use a demand and supply diagram to illustrate consumer and producer surplus.
   b. Identify the level of output where their sum is maximum.
   c. Explain the condition *MB* = *MC*, and use your diagram of part (a) to show at which level of output this condition is satisfied.
   d. Outline what the conditions of maximum consumer plus producer surplus, and *MB* = *MC* tell us about allocative efficiency, and what they tell us about social welfare.
4. Describe some limitations of the concept of maximum social welfare.
5. a. Using the diagram calculate consumer and producer surplus at the initial equilibrium given by S_1 and D.



   b. Suppose the supply curve shifts to S_2. Calculate the new consumer and producer surplus that arise at the new equilibrium.
   c. Calculate the increase or decrease in consumer and producer surplus due to the shift in the supply curve.
   d. Explain what accounts for the changes in consumer and producer surplus in terms of the changes in P and Q.

THEORY OF KNOWLEDGE 2.1

The meaning and implications of maximum social welfare

We must be very careful when interpreting the meaning of *maximum social welfare*. The achievement of *MB* = *MC*, or maximum social surplus, deals with the *what/how much to produce* and *how to produce* questions, and therefore with how a society can realise allocative efficiency.

Even if it is assumed that allocative efficiency is achieved, in which case we have 'maximum social welfare', this tells us nothing about how output and income are, or should be, distributed. The achievement of allocative efficiency is, in fact, consistent with any possible distribution of output or income. This means that we can have an extremely unequal distribution of output and income, where one person in society gets all the output, or a highly equal distribution where everyone gets an equal share, and both of these situations can lead to allocative efficiency. This idea highlights the point that *when we refer to maximum social welfare, we are only talking about making the best possible use of scarce resources, while saying nothing about who gets the benefits of what the resources produce*. For any distribution of income and output, ranging from the most unequal to the most equal, it is possible to have a situation of maximum social surplus and therefore allocative efficiency.

The pursuit of efficiency, dealing with the *what/how much to produce* and *how to produce* questions, is based on social scientific investigation, and tries to discover the most effective ways to increase efficiency in resource allocation. Many economists would argue that this is a matter of positive economics (see Chapter 1 on the distinction between positive and normative economics). On the other hand, the *for whom to produce* question is a matter of normative economics. Positive economic thinking, based on the scientific method described in Chapter 1, is not intended to make judgements about what is fair or unfair, or equitable or inequitable. It is intended to deal with issues of things that 'are' or 'will be' under different conditions. Therefore, while it can tell us about what methods are most likely to lead to increases in efficiency, it is not intended to tell us about how income and output should be distributed. The issue of equity (introduced in Chapter 1) is a normative issue, because what is considered 'fair' is a matter of beliefs and value judgements about things that 'ought to be'. What share of total income, or what particular goods and services *ought* individuals to have? Is there a minimum income that people *ought* to have? If so, based on what kinds of equity principles *ought* distribution and redistribution to take place?

There are no 'right' or 'wrong' answers to these questions. The answers that are chosen are based on beliefs and value judgements about what is good for society and the people within it. According to the standard view, economists should be concerned with positive aspects of economics. Their positive thinking could tell us about the likely consequences of policies to change the distribution of income, but should not make recommendations about how the distribution of income ought to change.

We will continue with the topic in Theory of knowledge 6.1 in Chapter 6.

Thinking points

- What is the significance of language in conveying meaning; does the expression 'maximum social welfare' accurately reflect its actual meaning?
- If economists, as social scientists, cannot make recommendations about normative issues like the distribution of income, how are these decisions made? (Think about the political process, social values, tradition and history.)
- Based on your reading about economics in the press and listening to the news, do you think that economists in the real world make a clear distinction between positive and normative ideas?
- Do you agree with the principle that economists (and social scientists generally) should only be concerned with positive thinking (social scientific investigation) and should leave normative issues (about things that ought to be) to societal decision-making?

2.6 Critique of the maximising behaviour of consumers and producers (HL only)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- evaluate the assumptions of consumer rationality, utility maximisation and perfect information (AO3)
- referring to the following points, discuss the limitations of standard consumer behaviour made by behavioural economists (AO3)
  - biases including rule of thumb, anchoring, framing, availability
  - bounded rationality
  - bounded self-control
  - bounded selfishness
  - imperfect information
- discuss policies inspired by behavioural economics (AO3)
  - choice architecture with respect to default, restricted and mandated choices
  - nudge theory
- distinguish between various objectives of businesses (AO3)
  - profit maximisation
  - alternative objectives including corporate social responsibility, market share, satisficing, growth

Rational economic decision-making

Standard economic theories and models are based on an important assumption, that of 'rational self-interest', or *rational economic decision-making*. This means that individuals are assumed to act in their best self-interest, trying to *maximise* (make as large as possible) the satisfaction they expect to receive from their economic decisions. It is assumed that consumers spend their money on purchases to maximise the satisfaction they get from buying different goods and services. Similarly, it is assumed that firms (or producers) try to maximise the profits they make from their businesses; workers try to secure the highest possible wage when they get a job; investors in the stock market try to get the highest possible returns on their investments, and so on.

As we will see in the discussion below, the assumption of rational economic decision-making and the principle of self-interested maximisation underlying economic theory has come to be seriously questioned.

Rational consumer choice

The standard theory of consumer behaviour is based on some assumptions of **rational consumer choice** which we discuss below.

Assumptions of rational consumer choice

Consumer rationality

According to *rational consumer choice* consumers make purchasing decisions according to their tastes and preferences which satisfy three assumptions.

1. The consumer is able to rank goods according to her preferences. This means that if confronted with two goods, or two groups of goods, A and B, the consumer can say with certainty that she prefers A to B, or she prefers B to A, or she is indifferent between A and B. (This is called the *completeness* assumption.)
2. Preferences among alternative choices are consistent. This means that if the consumer prefers A to B, and B to C, then she must prefer A to C. (This is called the *transitivity* assumption.)
3. The consumer always prefers more of a good to less. If there are two groups of goods where one group contains more of one good than the other group, the consumer will always prefer the group that has more of the one good. (This is called the *non-satiation* assumption.)

Perfect information

It is further assumed that consumers have at their disposal perfect information about all their alternatives, so that there is no uncertainty. In other words, when a consumer chooses good A over good B, she possesses all possible information about good A and good B, so that there are no doubts, questions or uncertainties about the goods being chosen. The consumer has knowledge of all possible products, product qualities and prices.

Utility maximisation

Utility, as we saw earlier in this chapter, refers to the satisfaction that consumers derive from consuming something. According to the theory of consumer behaviour, consumers *maximise their utility*, meaning they make it as large as possible, by buying the combination of goods and services that results in the greatest amount of utility for a given amount of money spent (the consumer's budget or income). This behaviour is based on the assumptions of rationality and perfect information that were presented above.

The concept of marginal utility, and the law of diminishing marginal utility, are fundamental building blocks of the theory of consumer behaviour that allow economists to show how consumers make their choices among alternatives in order to maximise utility.[5]

Behavioural economics: limitations of the assumptions of rational behaviour

Behavioural economics was introduced in Chapter 1, Section 1.5. It is a new branch of economics strongly influenced mainly by psychology, but also by sociology and neuroscience, based on the idea that human behaviour is far more complex than consumer rationality assumes. While the assumption of rationality has always been questioned on the grounds that people very often do not behave as rationally as is presumed, it is only since about the beginning of the 21st century that behavioural economics began to flourish into a distinct branch of economics.

Behavioural economics criticises consumer rationality and the idea of utility maximisation on the following grounds.

Biases

A **bias** (also known as *cognitive bias*) is a term from psychology that refers to systematic errors in thinking or evaluating. Biases are departures from normal standards of thought or judgment. Biases that affect consumer choices include the following:

- **Rules of thumb. Rules of thumb** are simple guidelines based on experience and common sense, simplifying complicated decisions that would have to be based on the complex consideration of every possible choice. Examples of rules of thumb are: one portion of salad is equal to two handfuls; a cup of filter coffee has 80 mg of caffeine; one serving of fruit is the size of a fist. These examples show that rather than make complicated measurements and calculations people often rely on simple rules to simplify decisions.
- **Anchoring**. **Anchoring** involves the use of irrelevant information to make decisions, which often occurs due to its being the first piece of information that the consumer happens to come across. For example, you find a pair of jeans you want that costs $150, and then you find a similar one for $100 that you buy thinking it's a bargain, but then you discover you could have gotten the same thing for $50 somewhere else.
- **Framing. Framing** deals with how choices are presented to decision-makers (framed). For example, consumers prefer beef described as 80% lean rather than 20% fat. Yet the rational consumer would be indifferent between the two as the information is exactly the same, it is only presented (framed) differently. In another example, consumers could be willing to pay a higher price for a faded pair of jeans in a boutique than for the identical pair of jeans in a discount store. In the first example the purchase is framed by language; in the second it is framed by the seller's environment.
- **Availability**. **Availability** refers to information that is most recently available, which people tend to rely on more heavily, though there is no reason to expect that this information is any more reliable than other information that was available at an earlier time. This may be due at least partly to the fact that people remember recent events or information more readily than older events or information. For example, if a consumer has recently received a lot of advertising information about the superiority of one brand over other brands, she may be more likely to select that brand, having been influenced by the recent advertising.

[5] It is beyond the scope of this book to explain how the law of diminishing marginal utility is used to explain rational consumer in order to maximise utility.

The presence of biases that influence consumer choices has led to alternative concepts and ideas put forward that may be more accurate descriptions of consumer behaviour:

Bounded rationality

Bounded rationality is an idea developed by Herbert Simon, who received the Nobel Prize in economics on 1971 for his work on economic decision-making. Simon argued that people do not have an unlimited capacity to process information and that searching for information needed to maximise utility is itself a costly process. He therefore suggested the concept of bounded rationality to describe consumer behaviour. Bounded rationality is the idea that consumers are rational only within limits, as consumer rationality is limited by consumers' insufficient information, the costliness of obtaining information, and the limitations of the human mind to process large amounts of information. According to Simon, rather than maximise, consumers *satisfice*, meaning they seek a satisfactory outcome rather than an optimal (or best) one.

Bounded self-control

Bounded self-control is related to bounded rationality (see above). It refers to the idea that people in reality exercise self-control only within limits. This means they often do not have the self-control that would be required of them to make rational decisions. People may at times eat too much, spend too much, save too little or work too little; all this is inconsistent with the assumption of rational behaviour.

Bounded selfishness

Another related concept is that of **bounded selfishness**, which is the idea that people are selfish only within limits, and that the assumption of self-interested behaviour underlying the maximisation principle cannot explain the numerous accounts of selfless behaviour and willingness to contribute to the public good even at the cost of reduced personal welfare (these points are further discussed in Theory of knowledge 2.2 below).

Imperfect information

The theory of consumer behaviour as we have seen is based on the assumption that consumers have access to all the information needed to make fully informed decisions, regarding prices, products, product quality. If they did not have access to such complete information, their decisions and choices would not be optimal, or best. However, it is abundantly clear that consumers do not and cannot possibly have such full access to information. This means they are unable to maximise utility as they make choices based on faulty and incomplete information.

THEORY OF KNOWLEDGE 2.2

Altruism, perceptions of fairness and self-interested behaviour of rational economic decision-makers

Economists have long been aware of the presence of altruism and acts based on perceptions of fairness in the behaviour of economic decision-makers. For example, when consumers make anonymous donations to charity or volunteer their services through many hours of work for the public good, they are not motivated by expectations of monetary gain. Firms, too, often act in the public interest by not firing workers in times of recession so as not to increase unemployment and cause personal hardship, even though this cuts into their profits.

Economists sometimes argue that there is no real conflict between such altruistic behaviour and behaviour based on self-interest, because people engaging in altruistic and charitable acts gain satisfaction from their contributions to society, and this works to increase their utility. Selflessness and caring for others are thus interpreted to be consistent with rational self-interest and with utility maximisation. Therefore in this view consumers are behaving rationally even when they behave altruistically.

A serious problem with this argument is that if the assumption of utility maximisation is interpreted so broadly that every human act is consistent with it, it loses its usefulness as a theory of consumer behaviour. To put it simply, if it explains everything, by being consistent with all human actions, it explains nothing. In the words of Herbert Simon, 'Economic theory has treated economic gain as the primary human motive.

CONTINUED

An empirically grounded theory would assign comparable weight to other motives, including altruism and the organisational identifications associated with it.'[6] In other words, altruism should be treated separately, and not be made consistent with the assumption of rational self-interest through an all-inclusive interpretation of utility.

Simon's thinking brings to mind the ideas of Adam Smith, discussed in Chapter 1, Section 1.5. In all likelihood, Smith would be extremely surprised and disappointed to discover that his contributions to economics regarding the invisible hand of markets have been taken out of context to mean that the best results for society arise because people behave *only on the basis of self-interest*. As noted in Chapter 1, in his early book, *The Theory of Moral Sentiments*, Smith argued that people's selfishness is restrained by 'moral sentiments', which are the need for self-respect and the respect of others, arising from empathy (which he refers to as 'sympathy') and the desire for honour and justice. Other economists of the classical period, among whom was David Ricardo, agreed with Smith that people's motivations are complex, involving a combination of self-interest and other motives like empathy and concern for other people. It was only much later, in the late 19th and early 20th centuries, that economists simplified and reduced human motivations solely to the idea of utility maximisation in the name of rational self-interest. This gave rise to the term *homo economicus* (from the Latin word *homo* meaning man, and the Greek word *economicus* meaning economic) or what translates into *economic man*, a mythical being who goes about life rationally maximising his self-interest. As Gary Becker, a noted Nobel Prize winner (1992) stated decades later, 'self-interest is assumed to dominate all other motives'.

Since Simon's time, economists and other scientists have carried out thousands of experiments that provide ample evidence of the presence of altruism in human beings. It is believed that altruism may derive from culture, or from psychological development and socialisation, or from innate biological traits. (See also Theory of knowledge 5.1 in Chapter 5 on sustainability and rational self-interest.)

Thinking points

Homo economicus, an assumed being who is forever rationally maximising utility as a consumer (or forever rationally maximising profit as a producer), is a fundamental building block of microeconomic theory. This mythical being came into existence at a time when economists were trying to imitate the methods of the sciences, especially physics. Just as physics assumes the existence of tiny particles (like atoms) that behave in a certain predictable way, so consumers and firms became the 'particles' of economics that behave in certain predictable ways. Although atoms are not visible with the naked eye, physicists have developed tools to detect their existence and describe their behaviour.

- What tools do economists use to make assumptions about the existence and behaviour of homo economicus?
- In what ways are physicists' tools the same or different from economists' tools? Do you think physicists' and economists' tools are reliable to the same extent?

Behavioural economics in action

Behavioural economics studies how individuals and groups of individuals make decisions. Instead of relying on assumptions about human behaviour, as in the theory of consumer behaviour, behavioural economics relies extensively on experiments that try to observe and understand how people behave and react in a variety of situations.

Nudge theory

The word *nudge* means poking someone gently to get attention. As an economics term, a **nudge** has come to mean a method designed to influence consumers' choices in a predictable way, without offering financial incentives or imposing sanctions, and without limiting choice. As an economics term the word *nudge* was coined by two American economists, Richard Thaler and Cass Sunstein

[6] Herbert Simon, Altruism and Economics, *Eastern Economic Journal* Vol. 18, No. 1 (Winter, 1992), pp. 73–83

in their now classic book of 2008, *Nudge: Improving Decisions about Health, Wealth and Happiness*. Nudges try to influence people to behave in socially desirable ways, but without imposing any constraints on their behaviour, and by maintaining freedom of choice. Richard Thaler received the Nobel Prize in Economics in 2017 in part for his work on behavioural economics.

While behavioural economics was initiated in the United States, it has taken a strong hold over policy in many countries in the world, with some countries establishing nudge units to pursue and support this approach to a variety of public policy areas. It is not only governments but also the private sector that have undertaken nudge initiatives.

Examples of nudges and their results include the following:

- In the United Kingdom, tax payments of late taxpayers increased by 15% when they were told by the UK revenue service that most people in their area had already paid their taxes. This worked because the late taxpayers were made to feel they were the exception to what was the norm in their community.
- In a Danish municipality, the streetlights turn red when solar panels are no longer sufficient to power the light, making residents more aware of their electricity consumption.
- Putting healthy foods in more visible and accessible positions in stores has been shown to increase sales of such foods. In addition, colourful footsteps leading consumers to the locations of healthy foods has also been effective.
- When non-payers of vehicle taxes received personalised messages rather than warning letters, the number of payers doubled; when the message included a photo of the car in question, the number of payers tripled.
- When taxpayers received notifications from tax authorities in white envelopes instead of brown with their names handwritten rather than typed, there was a 16% increase in people paying their tax.

These examples indicate that people respond favourably when *nudged* to perform actions that they might not otherwise be inclined to do.

Behavioural patterns such as these are taken as evidence that borrowing from psychology, sociology and neuroscience in an attempt to understand why human beings behave the way they do can have important results in changing behaviour to encourage achievement of socially desirable results.

Behavioural economics has been more successful in the area of finance, where it has already strongly influenced economic thinking. Other areas where it is making some headway are development, health, law and economics, public economics, organisational economics and wage determination.

Choice architecture

Choice architecture is the design of particular ways or environments in which people make choices; it is based on the idea that consumers make decisions in a particular context and that choices of decision-makers are influenced by *how options are presented to them*. Choice architects are individuals or organisations that arrange the context in which choices are made. The term *choice architecture* was also coined by Richard Thaler and Cass Sunstein in their book on nudges noted above. *Framing*, which was explained earlier, is a very important part of choice architecture, as it defines the context in which choices are presented to decision-makers, thus trying to influence their decisions.

Choice architecture offers different kinds of choices that are described below.

- **Default choice** is a choice that is made by default, which means doing the option that results when one does not do anything. People often make choices by default due to habit or lack of interest in taking a deliberate action, even if doing nothing may not the best choice for them. Sometimes they feel more comfortable not having to make a choice. Therefore one way of inducing people to follow a particular course of action is to provide it as a default choice. See Real world focus 2.3 below for an example of the role of default choice in organ donation.
- **Restricted choice** is a choice that is limited by the government or other authority. People everywhere are subject to countless restrictions of all kinds, ranging from speed limits, voting age, and recycling regulations to smoking prohibitions and social norms like shaking hands. It is argued that restrictions such as these are necessary because people have too many choices available to them, and in the absence of more and better information, or due to poor judgement, they often make poor choices. Therefore choice architecture can take advantage of restrictions to encourage people to make choices with socially desirable outcomes.

REAL WORLD FOCUS 2.3

Choice architecture, nudging and organ donation

Figure 2.19: Organ donor card

In the area of organ donation, it is a challenge to increase the number of organ donors in many countries around the world. Many countries have a *default choice* for people to be either (a) donors, in which case they automatically become a donor if they do nothing, and would have to make a deliberate effort to register as a non-donor; or (b) non-donors, in which case they automatically become a non-donor and would have to register to become a donor. In countries where the default option is to be a donor, for example Austria, Belgium, France, Hungary, Poland, Portugal, and the United Kingdom, the participation rate in organ donation is very high, sometimes as much as 90% of the population or more. By contrast, in countries where the default option is to be a non-donor, for example Denmark, Germany, and the United States, the participation rate is much lower.

Organ donation can also be used to illustrate *mandated choice*. In the state of Illinois in the United States, in order to get a driver's licence renewed it is mandatory (compulsory) to answer a question on whether one wishes to be an organ donor. The result has given rise to a 60% rate of signing up to be a donor, compared to only 38% for the US national average where mandated choice does not exist.

The default choice that makes people donors by default therefore appears to be effective. However, nudges may in some cases backfire, giving rise to opposite results than the desired ones. In the Netherlands, a 2016 law made being a donor the default choice, in the hope that the number of donors would increase. Yet in a drive that took place shortly after the bill was passed, the surprising result was that there were *six times as many non-donors as donors*. A likely explanation for such an unexpected result may be that the Dutch population rebelled against the nudger's intentions. There had been a lot of press coverage of the government's decision to change the default choice, and it appears the Dutch people did not want the government to make choices for them.

Sources: *Dee Gill, How to spot a nudge gone rogue, UCLA Anderson Review, School of Management, 12 September 2018*

Applying your skills

1 According to the experiences of many countries, the default choice for organ donation appears to be succeeding in increasing the number of donors. Can you think of other areas where this type of default choice could be applied?

2 What is the lesson about nudging that can be learned from the Dutch experience?

3 Some observers argue that nudging, in spite of its potential benefits, also carries a risk of manipulation of the public's opinion in subtle ways that may not always be detectable. What does the Dutch experience say about this possible risk? (See also Question 4 under *Inquiry and reflection* at the end of this chapter.)

- **Mandated choice** is a choice between alternatives that is made mandatory (compulsory) by the government or other authority. It can be thought of as required choice. It is a free choice, but it is compulsory to make that free choice. See Real world focus 2.3 for an example of the role of default choice in organ donation.

Note that default choice, restricted choice and mandated choice are different types of *nudges* within *choice architecture* that are intended to work toward influencing people's choices in a direction held to be socially desirable.

The response of mainstream economists to the critique of utility theory

In response to the criticisms of utility theory and rational behaviour, many mainstream economists would argue that in spite of its imperfections, the theory of consumer behaviour based on the assumption of rationality does a reasonably good job in explaining and predicting how consumers behave under many circumstances. There is evidence that consumers do respond rationally to monetary factors. For example, an increase in the gasoline or petrol tax will most likely lead to a lower quantity of petrol demanded, as predicted by the downward sloping demand curve based on consumer rationality. Examples abound of tests that have empirically shown (not refuted) the existence of downward sloping demand curves, which are based on the theory of utility maximisation. So even if consumers do not behave rationally all the time according to the rules of rationality, on average, overall, their behaviour is consistent with the predictions of utility theory.

Moreover, some economists would argue that choice architecture and the use of nudges is manipulative, as it attempts to exercise control and influence over consumer choices. The idea of free choice is therefore weakened and becomes an illusion of free choice rather than actual free choice. What may be worse is that it often affects consumers on a sub-conscious level, making them do things that they might not choose to do if they were fully aware of the full implications of their choices.

In effect, choice architecture takes advantage of the incomplete information available to the consumer, and selectively provides information (such as through framing and choice architecture) to encourage consumers to act in a particular way.

For these reasons it is argued that there should be transparency and open discussion of nudges used for economic policy.

Evaluating behavioural economics and economic policy

Potential advantages

- Behavioural economics may be a relatively simple and low-cost way to influence people's behaviour to act in socially desirable ways.
- It has been used successfully in a number of areas, suggesting that the methods of choice architecture and nudging may have numerous possible applications in areas that are as yet unexplored.
- It offers consumers and citizens, generally, freedom of choice without forcing them to do anything or preventing them from doing anything, hence without restricting their choices.
- It may be able to overcome the weaknesses of the theory of consumer behaviour, which is not always able to explain the inconsistencies and seeming irrationality of actual consumer behaviour.
- Policies are based on principles of psychology, such as framing, many of which have been previously tested over many years.
- The development of policies is based on trials, indicating the use of a flexible trial-and-error method of discovering policy measures that can work in achieving desired results.

Potential disadvantages

- The body of knowledge being developed is not based on any understanding of human behaviour, and is therefore unable to lead to a systematic and unifying theory of how consumers behave with general applicability.
- The resulting unsystematic approach (different policies for different situations) may not be valid over time or across different income groups, social groups or cultures; this reduces the applicability of the policies being developed over time and across social, economic and cultural groups.
- There may be risks of using psychological principles to manipulate consumers, in much the same ways that advertising and marketing have been using similar principles over many years in order to

manipulate consumer tastes and preferences for the purpose of convincing consumers to act in ways that are not necessarily in their best interests (such as advertising for sugary and fatty food products that are unhealthy).

- There may be risks that behavioural policies may be used as substitutes for necessary but politically costly economic policies, such as the imposition of taxes on socially harmful goods known as *demerit goods* or goods leading to *negative externalities* (see Chapter 5).
- Traditional economic policies (for example indirect taxes, subsidies, to be discussed in Chapters 4 and 5) may be more effective.
- It may be a new form of government regulation, only camouflaged under the guise of freedom of choice, and therefore more dangerous as people may not be aware of its existence and its effects on their choices.
- Choice architecture and nudging may successfully affect people's choices, but these choices may not be a reflection of their true preferences.

TEST YOUR UNDERSTANDING 2.9

1 a Discuss what it means to be 'a rational consumer' in economics.

  b To what extent are you personally a rational consumer?

2 For each of the biases discussed in the text, explain why they could not occur according to the theory of consumer behaviour.

3 In the examples of nudges given in the text, identify the framing element in the nudge.

4 Distinguish between default, restricted and mandated choice. Explain how these can be used as nudges.

5 Distinguish between bounded rationality, bounded self-control and bounded selfishness. Provide an example for each of these. Explain why they are all inconsistent with rational consumer behaviour.

6 Discuss the potential advantages and disadvantages of behavioural economics.

Firm business objectives

Rational producer behaviour: profit maximisation

Standard economic theory of the firm assumes **rational producer behaviour** according to which firms are guided by the goal to maximise profit. Profit maximisation involves determining the level of output that the firm should produce to make profit as large as possible.

On a very general level, profit is equal to the total revenue earned by a firm minus the total costs incurred by the firm in the process of producing its output.

Suppose you have a pizza restaurant and you sell 400 pizzas a day at \$7 each. You have revenues of 7 x 400 = \$2800 per day. On the other hand the costs of buying the pizza ingredients, paying your workers and meeting all other expenses of running your restaurant amount to \$2650 per day. Your profit per day is

Profit = revenues – costs = 2800 – 2650 = \$150 per day.

The objective of profit maximisation is to make the difference between revenues and costs as large as possible so as to make profit as large as possible. We will come back to the topic of profit maximisation in Chapter 7 (at HL).

Alternative business objectives

Over the years, economists have developed many theories about firm behaviour. The following is a brief survey of some of the more important ones. The common theme of all of these is that profit maximisation may not be the overriding objective of firms, as in fact they may have other goals that may be more important.

Corporate social responsibility: ethical and environmental concerns

The self-interested behaviour of firms often leads to negative consequences for society. It is often the case that the well-being of firms is not consistent with the well-being of society. A prime example is the self-interested firm that pollutes the environment. (Such actions of firms will be examined in Chapter 5 under the topics of negative externalities and common pool resources.) In addition, firms can engage in actions that most consumers would consider to be ethically unacceptable, such as the practice in many developing countries of employing poorly paid children forced to work long hours, or employing labour that is forced to work under unhealthy or dangerous conditions. These situations may arise in countries where there is widespread poverty, and government legislation protecting the rights of children and workers is either non-existent or poorly enforced.

However, many firms increasingly recognise that the pursuit of self-interest need not necessarily conflict with ethical and environmentally responsible behaviour. A negative image of the firm held by workers and buyers of the product can cut deeply into the firm's revenues and profits by lowering worker productivity and the firm's sales. Further, socially irresponsible firm behaviour may lead to government regulation of the firm intended to minimise the negative consequences of the firm's actions for society, whereas socially responsible behaviour could instead result in less government regulation. Therefore, firms face incentives to display **corporate social responsibility** by engaging in socially beneficial activities. These can take many forms, including:

- avoidance of polluting activities
- engaging in environmentally sound practices
- support for human rights, such as avoiding exploitation of child labour and labour in general in less developed countries, or avoiding investments in countries with politically oppressive regimes
- art and athletics sponsorships
- donations to charities.

Many of these practices are the result of increased consumer awareness of social and environmental issues, growing consumer concern over ethical and environmental aspects of business practices, and even consumer activism that results in boycotts of offending firms. One indication of the influence and concern of consumers is the rapidly growing interest in investments in companies (through stock markets) that meet certain social, ethical and ecological criteria.

Economists used to think that ethical and environmentally responsible behaviour of firms would reduce their profits. This was based on considering only the cost aspect of profits; for example, firms using cheap child labour face lower costs, and hence will make higher profits than firms avoiding such practices. Yet profits depend not only on costs, but also on revenues that will fall if consumers avoid buying the products of offending firms.

A number of studies have attempted to measure the effects of socially responsible behaviour on the profits of firms. Does ethical and environmentally responsible behaviour lower or increase firms' profits? The results of these studies have been inconclusive. The behaviour of firms themselves, however, suggests that they often do not want to risk consumer displeasure.

Market share

Market share refers to the percentage of total sales in a market that is earned by a single firm. For example, if a product has total sales within a country of $50 million, and a firm sells $10 million worth of that product, then that firm has a 20% market share. A high market share means that the firm is enjoying large sales of its product or products, and is an indication of the product's popularity among buyers.

A large market share means that it is likely that the firm is achieving *economies of scale* (falling costs per unit of output as the firm grows, to be studied at HL in Chapter 7). Economies of scale allow a firm to increase its profitability.

In general, market share is so important that large companies such as multinational corporations (large firms that have productive facilities in more than one country) often measure their performance in terms of their market share in specific countries. Market share is an important indicator of performance because it allows the firm to monitor how well it is doing in relation to its competitors.

A firm that succeeds in maintaining its market share over time means that it is likely earning revenues that are growing at the same rate as the overall market. A growing market share is a sign that the firm is doing well in relation to its competitors. It is likely that its revenues are growing faster than the revenues of competitor firms. A falling market share on the other hand means that the firm is not doing as well as its competitors in terms of sales of its products.

In order to increase its market share, the firm may try to lower its prices, or introduce new or innovative products into the market, or use advertising. Each of these strategies may have the effect of lowering profits. Lower prices may reduce revenues (though not necessarily, this will be explored in Chapter 3). The introduction of new products involves increased costs on research and development, while advertising also involves increased firm spending to pay for the advertising. Possible lower prices with higher costs will cut into profits. However, as long as the firm remains profitable, and it continues to earn a satisfactory level of profits, the benefits of maintaining or increasing market share are likely to make the lower profits worthwhile.

Growth maximisation

Another possible objective is maximisation of growth rather than profits.[7] **Growth maximisation** can be attractive for the following reasons:

- A growing firm can achieve economies of scale (see Chapter 7) and lower its average costs, thus increasing its profitability.
- As a firm grows it can diversify into production of different products and markets and reduce its dependence on a single product or market.
- A larger firm has greater market power and increased ability to influence prices, thus again potentially increasing its profitability.
- A larger firm reduces its risks because it may be less affected in an economic downturn (a recession) and is less likely to be taken over (bought) by another firm.
- The objective of growth maximisation reconciles the interests of both owners and managers, because both groups have much to gain from a growing firm. Other maximisation objectives may pit firm owners against firm managers; for example, profit maximisation is favoured by owners while revenue maximisation is favoured by managers.

Revenue maximisation

In one theory of firm behaviour, it is argued that the separation of firm management from firm ownership, which increasingly dominates business organisation, has meant that firms' objectives have changed. Whereas profit maximisation may be the dominant motive of the traditional owner-managed firm, firm managers who are hired by the owners to perform management tasks may be more interested in increasing sales and maximising the revenues that arise from larger quantities sold. This goal of firms is referred to as *revenue maximisation*.[8] Increasing sales and maximising revenues may be more useful to a firm than profit maximisation for the following reasons:

- Sales can be identified and measured more easily over the short term than profits, and increased sales targets can be used to motivate employees.
- Rewards for managers and employees are often linked to increased sales rather than increased profits.
- It is often assumed that revenue from more sales will increase more rapidly than costs; if this is the case, profit (= total revenues – total costs) will also increase.
- Increased sales give rise to a feeling of success, whereas declining sales create a feeling of failure.

(*Note that revenue maximisation, while important, is not in the syllabus.*)

Satisficing

Many of the above objectives assume that the firm tries to maximise some variable, whether it is profit, market share, growth or revenue. Herbert Simon, a Nobel Prize-winning economist (also mentioned earlier in connection with the idea of bounded rationality as a critique of consumer rationality), has argued that the large modern enterprise cannot be looked upon as a single entity with a single maximising objective; instead it is composed of many separate groups within the firm, each with its own objectives which may overlap or may conflict. This multiplicity of objectives does not allow the firm to pursue any kind of maximising behaviour. Firms therefore try to establish processes through which they can make compromises and reconcile conflicts to arrive at agreements, the result of which is the pursuit of many objectives that are placed in a hierarchy. This behaviour was termed **satisficing** by Simon. (Simon used this term to describe the behaviour of consumers as well.) It refers to the idea that firms try to achieve a satisfactory level of profits together with satisfactory results for many more objectives, rather than optimal or 'best' results for any one objective.

TEST YOUR UNDERSTANDING 2.10

Imagine you are the owner of a business. Discuss and justify any business goals you may have that are not directly concerned with profit maximisation.

[7] This is based on the work of R. Marris and others.
[8] The revenue-maximisation goal of firms was described by W. J. Baumol in 1959.

THEORY OF KNOWLEDGE 2.3

How important are the criticisms of profit maximisation as the firm's main goal?

Standard economic theory assumes that profit maximisation is the most important goal of firms. As we will see in Chapter 7, the study of firm behaviour is based very heavily on the assumption of profit maximisation. Yet this assumption is criticised for several reasons:

- The use of marginal concepts (for example *marginal cost*) in the theory is unrealistic; firms cannot easily identify marginal costs, and do not even try to do so; therefore, this theory does not accurately describe methods actually used by firms to determine price and output.
- The model is based on the assumption that firms have perfect information at their disposal, whereas in fact the information on which they base their decisions is highly fragmentary and uncertain; firms do not know what demand curves they face for their products and they do not know how competitor firms will behave in response to their actions.
- Short-run profit maximisation may be unrealistic; firms may prefer lower profits in the short run in exchange for larger profits over the long run.
- The factors determining demand and supply for products and resources are continuously changing, with demand and supply curves continuously shifting, so that any profit-maximising decisions regarding prices and output made today under current conditions may be irrelevant by the time the output is produced and ready for sale in the market.
- There is real-world evidence suggesting that firm behaviour may be motivated by a variety of objectives other than profit maximisation, which were discussed above.

Milton Friedman, an American Nobel Prize-winning economist, argued in a famous book[9] that it does not matter if the assumptions of a theory are unrealistic, as long as the theory has predictive powers. In fact, good theories are often based on unrealistic assumptions that do not accurately describe the real world, because the role of assumptions is to portray only the important aspects of a process that is modelled or theorised about, ignoring the irrelevant details.

Paul Samuelson, another American Nobel Prize-winning economist, fundamentally disagreed. Samuelson argued that the predictions of a theory can only be as empirically valid as the theory itself, and as the assumptions on which the theory rests. If the assumptions are unrealistic or invalid, then the theory and its predictions will similarly be invalid; it is not possible to have a theory with predictive powers if its assumptions are unrealistic. If the predictions of a theory are empirically valid, so is the theory and its assumptions. Logically, then, it is not possible to separate the predictions of a theory from the assumptions of the theory; they all stand or fall together.

Thinking points

- Remember that a theory tries to explain real-world events. Does it matter if a theory is based on unrealistic assumptions?
- As you read through Chapter 7, you may want to keep these issues in mind, as we will encounter further unrealistic assumptions in some market models discussed in that chapter (see also Theory of knowledge 7.1 about perfect competition).

[9] Milton Friedman (1953) 'The Methodology of Positive Economics' in Essays in Positive Economics, University of Chicago Press.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1. Advertising makes great use of biases that influence consumers' way of thinking. As you watch TV, when a commercial comes on, try to see if the message being conveyed to the consumer makes use of biases in order to convince consumers to buy the product.
2. An area that has been examined extensively involves providing healthy lunches at schools and using nudges to encourage parents and students to prefer these. Investigate what kinds of nudges have been suggested, when and where they have been implemented, and what were the results. Try to design a programme where you could recommend nudges that you think may be effective in increasing demand for healthy school meals.
3. Was your decision to study IB economics a rational choice? Identify the conditions that would have to be satisfied for this choice to have been a rational one. If it was not rational, identify some biases that may have affected your choice.
4. Richard Thaler and Cass Sunstein, who became famous for their contributions to nudges. coined the phrase *libertarian paternalism* to describe nudges as a form of government intervention. The word *libertarian* (from the Latin word *libertas* which means freedom) refers to a philosophy where liberty or freedom and autonomy are the main principles, meaning that people should have freedom of choice and should be free of authority. The word *paternalism* (from the Greek word *pater* which means father) refers to actions that limit people's freedom or autonomy; such actions are often justified on the grounds that they are in the people's best interests. Consider the phrase *libertarian paternalism*, and determine whether these two concepts may be contradictory or whether they can be made consistent with each other as Thaler and Sunstein have tried to do through the policy of nudges.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 3

Elasticities

BEFORE YOU START

- When the price of a good or service that you want rises, how likely is it that you will purchase it anyway? You might say it depends. What factors might influence your decision?
- If producers of a good or service want to increase their revenue, sometimes they increase their price. Does this strategy always work? Why or why not?
- If someone is poor and over time becomes richer, what kinds of changes might you expect to see in her or his purchasing habits? What if she or he is rich and over time becomes poorer?

The topics above relate to the concept of elasticity. **Elasticity** is a measure of the responsiveness of a variable to changes in price or any of the variable's determinants. In this chapter we will examine three kinds of elasticities:

- price elasticity of demand, which examines the responsiveness of quantity demanded to changes in price
- income elasticity of demand, which examines the responsiveness of demand to changes in income
- price elasticity of supply, which examines the responsiveness of quantity supplied to changes in price.

Each of these has a number of applications to important economic problems.

Photo: Group of people shopping

3.1 Price elasticity of demand (*PED*)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- use the formula for price elasticity of demand (*PED*) to calculate *PED*, changes in price, changes in quantity and total revenue (AO4)
- identify the various degrees and range of values for *PED* (AO2)
- draw diagrams showing the range of values for *PED*, including (AO4)
  - relatively elastic and inelastic demand
  - constant values for perfectly elastic demand, perfectly inelastic demand and unitary *PED*
- explain and illustrate in a diagram changing *PED* along a straight-line downward sloping demand curve (HL only) (AO2, AO4)
- analyse the determinants of *PED* (AO2)
- explain the relationship between *PED* and total revenue (AO2)
- draw a diagram to show how total revenue changes in response to a price change depending on elastic or inelastic demand (AO4)
- apply *PED* to:
  - discuss its importance for firms and government decision-making (AO3)
  - analyse the reasons why primary commodities generally have a lower *PED* than manufactured products (HL only) (AO2)

Price elasticity of demand

Understanding price elasticity of demand (*PED*)

According to the law of demand, there is a negative relationship between price and quantity demanded: the higher the price, the lower the quantity demanded, and vice versa, *ceteris paribus*. We now want to know *by how much* quantity responds to change in price.

Price elasticity of demand (*PED*) is a measure of the responsiveness of the quantity of a good demanded to changes in its price. *PED* is calculated along a given demand curve. In general, if quantity demanded is highly responsive to a change in price, demand is referred to as being *price elastic*; if quantity demanded is not very responsive, demand is *price inelastic*.

The formula for *PED*

Suppose we are considering *PED* for good *X*. The formula used to measure its *PED* is:

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

If we abbreviate 'change in' by the Greek letter Δ, this formula can be rewritten as:

$$PED = \frac{\%\Delta Q}{\%\Delta P}$$

Simplifying, the above formula can be rewritten as:

$$\text{PED} = \frac{\frac{\Delta Q}{Q}\times 100}{\frac{\Delta P}{P}\times 100} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} \text{ where}$$

$$\Delta Q = Q_{\text{final value}} - Q_{\text{initial value}}$$

$$Q = Q_{\text{initial value}}$$

$$\Delta P = P_{\text{final value}} - P_{\text{initial value}}$$

$$P = P_{\text{initial value}}$$

The sign of *PED*

Since price and quantity demanded are negatively related, *PED* is a negative number. Any percentage increase in price (a positive denominator) gives rise to a percentage decrease in quantity demanded (a negative numerator), leading to a negative *PED*. Similarly, a percentage price decrease results in a percentage quantity increase. However, *the common practice is to drop the minus sign and consider PED as a positive number*. (In mathematics this is called taking the absolute value.) This is done to avoid confusion when making comparisons between different values of *PED*. Using positive numbers, we can say, for example, that a *PED* of 3 is larger than a *PED* of 2. (Had we been using the minus sign, −2 would be larger than −3.)

The use of percentages

Elasticity is measured in terms of percentage changes of P and Q for two reasons:

- We need a measure of responsiveness that is independent of units. It makes little sense to compare units of oranges with units of computers or cars. Also, we want to be able to compare elasticities across countries that have different currencies; an elasticity measured in terms of euros will not be comparable with an elasticity measured in yen or pounds. The use of percentages allows us to express elasticities in common terms.
- It is meaningless to think of changes in prices or quantities in absolute terms (for example, a \$15 increase in price or a 20 unit decrease in quantity). A \$15 price increase means something very different for a good whose original price is \$100 than for a good whose original price is \$5000. In the first case there is a 15% increase, and in the second there is a 0.3% increase. Using percentages allows us to put responsiveness into perspective.

The same arguments apply to all other elasticities we will consider.

Calculating *PED*, change in price and change in quantity

We can now use the formula above to calculate *PED*. Suppose consumers buy 6000 TVs when the price is \$255 per unit, and they buy 5000 TVs when the price is \$300.

$$PED = \frac{\frac{6000-5000}{5000}}{\frac{255-300}{300}} = \frac{\frac{1000}{5000}}{\frac{-45}{300}} = \frac{0.20}{-0.15} = -1.33$$

or 1.33 since we drop the minus sign. Therefore *PED* for TVs is 1.3.[1]

Note that elasticity is measured as a number, not as a percentage, and there are no units.

If we know *PED* and the percentage change in P, we can calculate the percentage change in Q. Suppose *PED* = 1.25 and price of good X *increases* by 12%. Calculate the percentage change in Q demanded:

$PED = 1.25 = \dfrac{\%\Delta Q}{-0.12} \Rightarrow \%\Delta Q = -0.12x1.25 = -0.15$ or −15%. Quantity *decreased* by 15%.

Similarly, if we know *PED* and the percentage change in Q demanded, we can calculate the percentage change in P.

[1] You may note that the value of this elasticity of demand depends on the choice of the initial price–quantity combination. In the calculation above, this was taken to be 300, 5000. If we had taken 255, 6000 as the initial price–quantity combination, we would get a *PED* value of 0.94. (You could calculate this as an exercise.) This difficulty can be overcome by use of the 'midpoint formula'. Note that you are not required to know the midpoint formula:

$$PED = \frac{\frac{\Delta Q_X}{\text{average}\,Q_X}}{\frac{\Delta P_X}{\text{average}\,P_X}}$$

In the previous example,

$$PED = \frac{\frac{1000}{5500}}{\frac{45}{277.5}} = 1.12, \text{ where } 5500 = \frac{(5000+6000)}{2} \text{ and } 277.5 = \frac{(255+300)}{2}$$

i.e. we use the average of the two Q_x values and the average of the two P_x values instead of the initial Q_x and initial P_x. Note that you are not required to know the midpoint formula.

Suppose $PED = 0.80$ and quantity of good Y demanded *falls* by 16%. Calculate the percentage change in P:

$$PED = 0.80 = \frac{-0.16}{\%\Delta P} = \Longrightarrow 0.80 \times \%\Delta P = -0.16$$

$\Rightarrow \%\Delta P = 0.20$ or 20%. Price *increased* by 20%.

You may note that here we have used negative signs in the calculations to show decreases, however this is not necessary since the final elasticity value is taken as a positive number (absolute value).

Total revenue calculations (referred to in the learning objectives) will be presented below.

TEST YOUR UNDERSTANDING 3.1

1 a Explain the meaning of price elasticity of demand.

  b Outline why we say it measures responsiveness of quantity along a given demand curve.

2 State why we treat *PED* as if it were positive, even though it is usually negative.

3 It is observed that when the price of pizzas is \$16 per pizza, 100 pizzas are sold; when the price falls to \$12 per pizza, 120 pizzas are sold. Calculate price elasticity of demand.

4 A 10% increase in the price of a particular good gives rise to an 8% decrease in quantity bought. Calculate the price elasticity of demand.

5 The *PED* for good X is 0.8. If the price of good X increased by 15%, calculate the percentage decrease in quantity demanded.

6 The *PED* of good Y is 1.5. If quantity of good Y increases by 30%, calculate the percentage decrease in the price of good Y.

The range of values for *PED*

The value of *PED* involves a comparison of two numbers: the percentage change in quantity demanded (the numerator in the *PED* formula) and the percentage change in price (the denominator). This comparison yields several possible values and range of values for *PED*. These are illustrated in Figure 3.1 and summarised in Table 3.1.

- **Demand is price inelastic when *PED* < 1 (but greater than zero)**. The percentage change in quantity demanded is smaller than the percentage change in price, so the value of *PED* is less than one; quantity demanded is relatively unresponsive to changes in price, and demand is **price inelastic**. Figure 3.1(a) illustrates price inelastic demand: the percentage change in quantity demanded (a 5% decrease) is smaller than the percentage change in price (a 10% increase), therefore *PED* is less than one.
- **Demand is price elastic when *PED* > 1 (but less than infinity**). The percentage change in quantity demanded is larger than the percentage change in price, so the value of *PED* is greater than one; quantity demanded is relatively responsive to price changes, and demand is **price elastic**. In Figure 3.1(b) the percentage change in quantity demanded (−10%) is larger than the percentage change in price (5%), therefore *PED* is greater than one.

In addition, there are three special cases where PED is constant (unchanging) along the full length of the demand curve:

- **Demand is unit elastic when *PED* = 1**. The percentage change in quantity demanded is equal to the percentage change in price, so *PED* is equal to one; demand is then *unit elastic*; there is **unitary *PED***. Figure 3.1(c) shows a unit elastic demand curve, where the percentage change in quantity demanded (−5%) is equal to the percentage change in price (5%).
- **Demand is perfectly inelastic when *PED* = 0**. The percentage change in quantity demanded is zero; there is no change in quantity demanded, which remains constant at Q_1 no matter what happens to price; *PED* is then equal to zero and demand is **perfectly inelastic**. For example, a heroin addict's quantity of heroin demanded is unresponsive to changes in the price of heroin. Figure 3.1(d) shows that a perfectly inelastic demand curve is vertical.
- **Demand is perfectly elastic when *PED* = infinity.** When a change in price results in an infinitely large response in quantity demanded, demand is **perfectly elastic**. As shown in Figure 3.1(e) the perfectly elastic demand curve is horizontal. At price P_1, consumers will buy any quantity that is available. If price falls, buyers will buy all they can (an infinitely large response); if there is an increase in price, quantity demanded drops to zero. This apparently strange kind of demand will be considered in Chapter 7 (at HL).

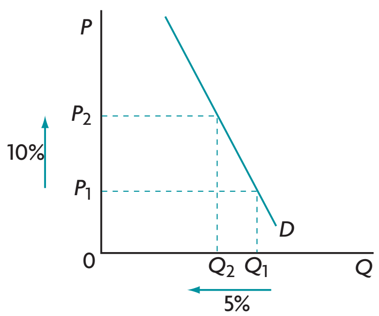
The numerical value of *PED* can therefore vary from zero to infinity. In general, the larger the value of *PED*, the greater the responsiveness of quantity demanded. While the cases of unit elastic, perfectly inelastic and perfectly elastic demand are rarely encountered in practice, they have important applications in economic theory.

Value of *PED*	Classification	Interpretation
Frequently encountered cases		
$0 < PED < 1$ (greater than zero and less than one)	price inelastic demand	quantity demanded is relatively unresponsive to price
$1 < PED < \infty$ (greater than 1 and less than infinity)	price elastic demand	quantity demanded is relatively responsive to price
Special cases: constant PED along the length of the demand curve		
$PED = 1$	unit elastic demand	percentage change in quantity demanded equals percentage change in price
$PED = 0$	perfectly inelastic demand	quantity demanded is completely unresponsive to price
$PED = \infty$	perfectly elastic demand	quantity demanded is infinitely responsive to price

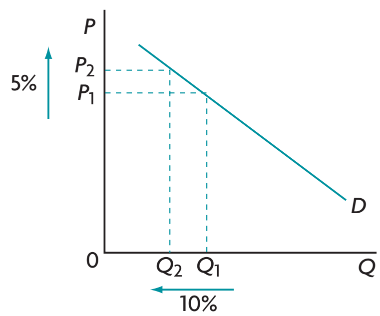
Table 3.1: Characteristics of price elasticity of demand

Frequently encountered cases

a Price inelastic demand: $0 < PED < 1$

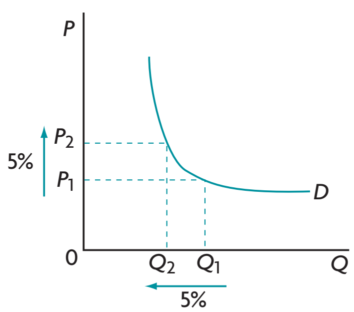


b Price elastic demand: $1 < PED < \infty$

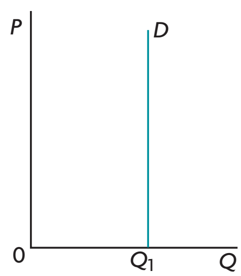


Special cases

c Unit elastic demand: $PED = 1$



d Perfectly inelastic demand: $PED = 0$



e Pefectly elastic demand: $PED = \infty$

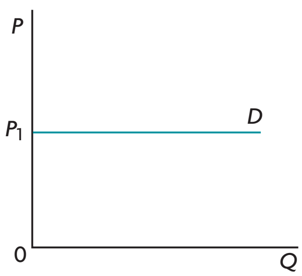


Figure 3.1: Demand curves and *PED*

PED and the steepness of the demand curve

The variety of demand curves and their *PED*s in Figure 3.1 suggest that the flatter the demand curve, the more elastic the demand (the higher the *PED*); the steeper the demand curve, the less elastic the demand (the lower the *PED*).

However, it is not always accurate to conclude that demand is more elastic or less elastic in different demand curves simply by comparing their steepness. The reason is that demand curves drawn on different scales are not comparable. Figure 3.2 shows two identical demand curves with different scales on the horizontal axis. It would be incorrect to conclude that the steeper demand curve has a less elastic demand.

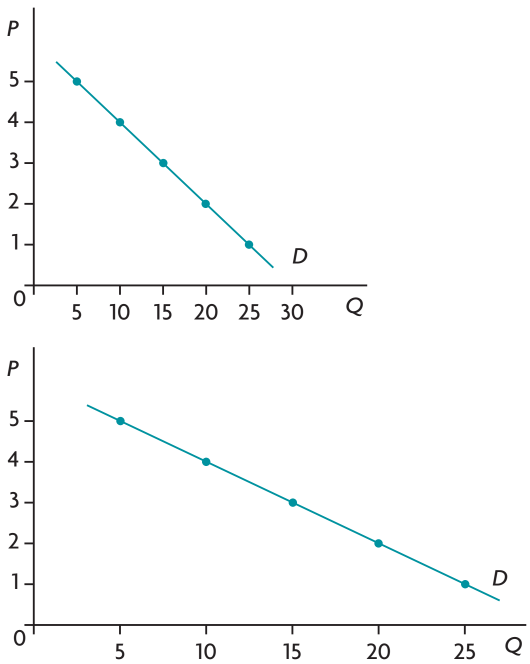


Figure 3.2: Two identical demand curves

So when is it correct to compare *PED*s of demand curves by referring to their steepness? *This can be done when* the demand curves intersect at some point, such as demand curves D_1 and D_2 in Figure 3.3. In this figure, for any price, D_1 is flatter and more elastic than D_2. For example, if price falls from P_1 to P_2, the resulting percentage change in quantity will be larger for D_1 (increase from Q_1 to Q_3) than for D_2 (increase from Q_1 to Q_2). In general, when demand curves intersect, then for any given price, the flatter the demand curve, the more elastic is the demand. This generalisation holds for comparisons between two demand curves at a particular price.

We often use the relative steepness of demand curves to be an indication of *PED*, so that comparing two demand curves, the one that is flatter is said to be more elastic while the one that is steeper is said to be more inelastic. (This is done on the assumption that if they were drawn on the same diagram they would intersect at some point.)

Intersecting demand curves and PEDs

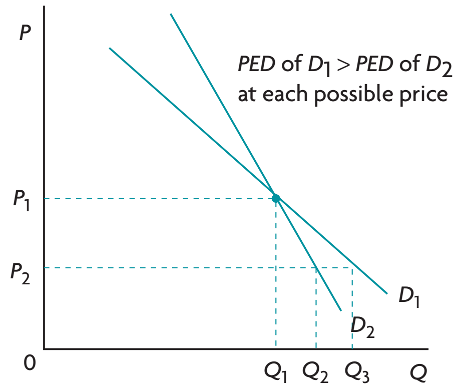


Figure 3.3: Demand curves and PEDs

TEST YOUR UNDERSTANDING 3.2

1 Specify the value for each of the following *PED*s and show, using diagrams, the shape of the demand curve that corresponds to each one:

  a perfectly elastic demand,

  b unitary *PED* (unit elastic demand) and

  c perfectly inelastic demand.

2 Provide examples of goods likely to have demand that is

  a elastic, and

  b inelastic.

3 Identify which price elasticity of demand values or range of values we see most frequently in the real world.

4 Assuming two demand curves intersect, explain which of the two will be relatively more elastic for a given price.

Changing *PED* and the straight-line demand curve (HL only)

Why *PED* varies

Along any *downward-sloping, straight-line demand curve*, the *PED* varies (changes) as we move along the curve. This applies to all demand curves of the types shown in Figure 3.1(a) and (b). It excludes unit elastic, perfectly inelastic and perfectly elastic demand curves (where *PED* is constant). We can see in Figure 3.4 that when price is low and quantity is high, demand is inelastic; as we move up the demand curve towards higher prices and lower quantities, demand becomes more and more elastic. The figure shows the *PED* values along different parts of the demand curve.

The reason behind the changing *PED* along a straight-line demand curve has to do with how *PED* is calculated. At high prices and low quantities, the percentage change in Q is relatively large (since the denominator of $\Delta Q/Q$ is small), while the percentage change in P is relatively small (because the denominator of $\Delta P/P$ is large). Therefore, the value of *PED*, given by a large percentage change in Q divided by a small percentage change in P results in a large *PED* (elastic demand). At low prices and high quantities, the opposite holds. The value of *PED* is given by a low percentage change in Q divided by a high percentage change in P, resulting in a low *PED* (inelastic demand).

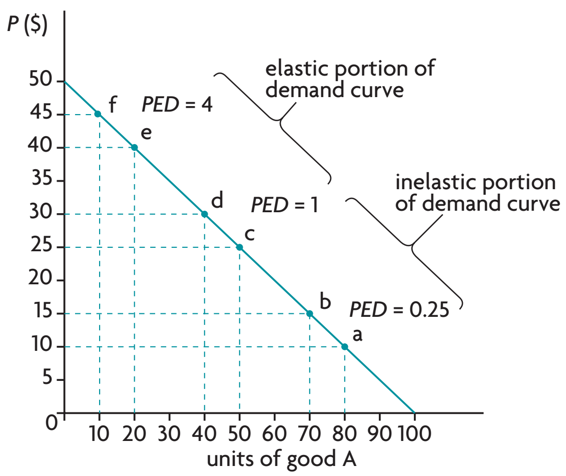


Figure 3.4: Variability of *PED* along a straight-line demand curve

On any downward-sloping, straight-line demand curve, demand is price-elastic at high prices and low quantities, and price-inelastic at low price and large quantities. At the midpoint of the demand curve, there is unit elastic demand.

Therefore, the terms 'elastic' and 'inelastic' should not be used to refer to an entire demand curve (with the exception of the three special cases where *PED* is constant), but only to a portion of the demand curve corresponding to a particular price or price range.

Why *PED* varies along a straight-line demand curve from a different perspective

A far simpler explanation of the varying *PED* along a straight-line demand curve is the following.

Note that the slope of demand curve[2] $= \dfrac{\Delta P}{\Delta Q}$, while

$$PED = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\Delta Q}{\Delta P} \times \frac{P}{Q} = \frac{1}{slope} \times \frac{P}{Q}$$

Now we know that the slope along a straight line is always constant, therefore the inverse of the slope = 1/slope is also constant.

Therefore *PED* is equal to a constant number multiplied by P/Q. As we move down the demand curve, we know that P gets smaller and smaller, while Q gets bigger and bigger, therefore P/Q becomes continuously smaller. This means that as we move down the demand curve, *PED*, which is equal to a constant number times an ever decreasing number, must be getting smaller and smaller. We thus have a simple explanation of why *PED* continuously decreases as we move down the demand curve.

[2] It may be noted that strictly speaking the slope of the demand curve is ΔQ/ΔP. The reason is that mathematically the slope is defined as the change in the dependent variable divided by the change in the independent variable. In the case of demand, Q is the dependent variable and P is the independent variable. However because of the practice of reversing the P and Q axes, and putting P on the vertical axis, we may take the slope as being ΔP/ΔQ. In any case this does not make any difference to the point of the argument in the text because both ΔQ/ΔP and ΔP/ΔQ are constant along a straight line.

TEST YOUR UNDERSTANDING 3.3

1 Using the information in Figure 3.4, calculate *PED* between:

a points a and b, where price increases from $10 to $15;

b points c and d, where price increases from $25 to $30; and

c points e and f, where price increases from $40 to $45.

d State the general principle about values of the *PED* along the straight-line demand curve that your calculations show.

Determinants of price elasticity of demand

We will now consider the factors that determine whether the demand for a good is elastic or inelastic.

Number and closeness of substitutes

The more substitutes a good (or service) has, the more elastic is its demand. If the price of a good with many substitutes increases, consumers can switch to other substitute products, therefore resulting in a relatively large drop (large responsiveness) in quantity demanded. For example, there are many brands of toothpaste, which are close substitutes for each other. An increase in the price of one, with the prices of others constant will lead consumers to switch to the others; hence demand for a specific toothpaste brand is price elastic. If a good or service has few or no substitutes, then an increase in price will bring forth a relatively small drop in quantity demanded. An increase in the price of petrol (gasoline) is likely to lead to a relatively small decrease in quantity demanded, because there are no close substitutes; therefore, demand for petrol is price inelastic.

Also important is the closeness of substitutes. For example, Coca-Cola® and Pepsi® are much closer substitutes than Coca-Cola and orange juice; we say that Coca-Cola and Pepsi have greater *substitutability*. The closer two substitutes are to each other, the greater the responsiveness of quantity demanded to a change in the price of the substitute, hence the greater the *PED*, because it is easier for the consumer to switch from one product to the other.

A factor that affects the number of substitutes is whether the good is defined broadly or narrowly. For example, *fruit* is a broad definition in comparison to *specific fruits* like oranges, apples and pears, which are narrowly defined. If we had considered fruit in relation to food, *food* is broad and *fruit* is narrow. The point here is that the narrower the definition of a good, the greater the number of close substitutes and the more elastic the demand. The demand for apples is more elastic than the demand for fruit; demand for fruit is more elastic than the demand for food. Similarly, a Honda has a higher *PED* than all cars considered together.

Necessities versus luxuries

Necessities are goods or services we consider to be essential or necessary in our lives; we cannot do without them. **Luxuries** are not necessary or essential. The demand for necessities is less elastic than the demand for luxuries. For example, the demand for medications tends to be very inelastic because people's health or life depend on them; therefore, quantity demanded is not very responsive to changes in price. The demand for food is also inelastic, because people cannot live without it. On the other hand, the demand for diamond rings is elastic as most people view them as luxuries. In general, the more necessary a good, the less elastic the demand.

A special case of necessity is a consumer's addiction to a good. The greater the degree of addiction to a substance (alcohol, cigarettes, and so on), the more inelastic is the demand. A price increase will not bring forth a significant reduction in quantity demanded if one is severely addicted.

Length of time

The longer the time period in which a consumer makes a purchasing decision, the more elastic the demand. As time goes by, consumers have the opportunity to consider whether they really want the good, and to get information on the availability of alternatives to the good in question. For example, if there is an increase in the price of heating oil, consumers can do little to switch to other forms of heating in a short period of time, and therefore demand for heating oil tends to be inelastic over short periods. But as time goes by, they can switch to other heating systems, such as gas, or they can install better insulation, and demand for heating oil becomes more elastic.

Proportion of income spent on a good

The larger the proportion of one's income needed to buy a good, the more elastic the demand. An item such as a pen takes up a very small proportion of one's income, whereas a car takes up a much larger proportion. For the same percentage increase in the price of pens and in the price of cars, the response in quantity demanded is likely to be greater in the case of the car than in the case of pens.

TEST YOUR UNDERSTANDING 3.4

1 Identify and explain the determinants of price elasticity of demand.

2 State in which case demand is likely to be more elastic in each of the following pairs of goods, and why:

- a chocolate or Cadbury's chocolate
- b orange juice or water
- c cigarettes or sweets

CONTINUED

- d a notepad or a computer
- e heating oil in one week or in one year
- f bread or meat.

Applications of price elasticity of demand

Price elasticity of demand is a very important concept in economics, with numerous applications. Some of these will be considered below; others will be studied in later chapters.

PED and total revenue

PED and the effects of price changes on total revenue

Total revenue (*TR*) is the amount of money received by firms when they sell a good (or service), and is equal to the price (*P*) of the good times the quantity (*Q*) of the good sold. Therefore, $TR = P \times Q$.

REAL WORLD FOCUS 3.1

What happens when demand is highly price inelastic?

A girl sells lemonade at a stand for 50 cents (= $0.50) a cup. On a very hot day, the lemonade becomes even more popular, and the girl realises she can raise her price a little and still sell all her lemonade. One afternoon, a diabetic boy comes along asking for lemonade with extra sugar because his blood sugar has fallen to dangerously low levels. The girl sees an opportunity and increases her price by 500%. The boy doesn't have enough money, but she tells him she will give him the lemonade right away provided he promises to run home afterward, get the money and return to pay her the full price. Having no choice, the boy agrees.

Figure 3.5: Little girl selling lemonade

Applying your skills

1 a Describe the boy's price elasticity of demand for sweet lemonade at that particular moment.

b Identify what determinant of *PED* accounts for this.

2 What would have happened to the quantity of lemonade demanded if the other children were faced with a 500% increase in its price? Explain in terms of their price elasticity of demand for lemonade.

Source: *Adapted from Teymour Semnani, 'Free markets don't always do the right thing regarding health care' in The Deseret News, 15 November 2009.*

We are interested in examining what will happen to the firm's total revenue (*TR*) when there is a change in the price of the good it produces and sells. We know that an increase in *P* leads to a decrease in *Q* demanded and vice versa. What can we say about the resulting change in total revenue? Will it increase, decrease or stay the same? The change will depend on price elasticity of demand of the good. We have the following three possibilities:

Demand is elastic (*PED* > 1)

When demand is elastic, an increase in price causes a fall in total revenue, while a decrease in price causes a rise in total revenue. To see why, consider that if demand is elastic, so *PED*>1, an increase in price results in a proportionately larger decrease in quantity demanded. For example, if price rises by 10% quantity demanded will fall by more than 10%. The decrease in quantity has a bigger impact on total revenue than the increase in price; therefore, total revenue falls. If there is a price decrease, a 10% price fall results in a larger than 10% increase in quantity demanded, and total revenue increases.

> When demand is elastic, an increase in price causes a fall in total revenue, while a decrease in price causes a rise in total revenue.

Demand is inelastic (*PED* < 1)

When demand is inelastic, an increase in price causes an increase in total revenue, while a decrease in price causes a fall in total revenue. Since *PED* < 1, an increase in price causes a proportionately smaller decrease in quantity demanded. For example, a 10% price increase produces a smaller than 10% decrease in quantity demanded, and total revenue rises. If price falls, a percentage price decrease gives rise to a smaller percentage increase in quantity demanded and total revenue falls. In both cases, the effect on total revenue of the change in price is larger than the effect of the change in quantity.

> When demand is inelastic, an increase in price causes an increase in total revenue, while a decrease in price causes a fall in total revenue.

Demand is unit elastic (unitary *PED*; *PED* = 1)

When demand is unit elastic, the percentage change in quantity is equal to the percentage change in price, and total revenue remains constant.

> When demand is unit elastic, a change in price does not cause any change in total revenue.

Using diagrams to illustrate *PED* and the effects of price changes on total revenue

Figure 3.6 shows three demand curves. The first shows a price range where *PED*>1, the second a price range where *PED*<1 and the third unitary *PED* throughout its range. In all three diagrams total revenue or *TR* is shown by the rectangles represented by $P \times Q$.

Elastic demand: as P increases TR falls: P and TR change in opposite directions

In Figure 3.6(a), where demand is elastic, at the initial price and quantity, P_1 and Q_1, total revenue is given by the sum of the rectangles A and B. When price increases to P_2 and quantity drops to Q_2, total revenue is given by the sum of the rectangles A and C. Due to the price increase rectangle B was lost and the rectangle C was gained. Since the loss (B) is larger than the gain (C), total revenue fell.

We can use the same diagram to explore a price decrease when *PED* > 1, simply by assuming that the initial price and quantity are P_2 and Q_2; price then falls to P_1 while quantity increases to Q_1. The gain in *TR* is given by rectangle B, which is greater than the loss shown by rectangle C, thus total revenue increases.

Inelastic demand: as P increases TR increases: P and TR change in same direction

These results can be seen in Figure 3.6(b). With a price increase, total revenue gained (rectangle C) is larger than total revenue lost (rectangle B); therefore, total revenue increases. If price falls from P_2 to P_1, the gain in total revenue (rectangle B) is smaller than the loss (rectangle C) and total revenue falls.

Unitary *PED*: any change in P leaves TR unchanged

In Figure 3.6(c), as price and quantity change, the gain in total revenue is exactly matched by the loss, and total revenue remains unchanged.

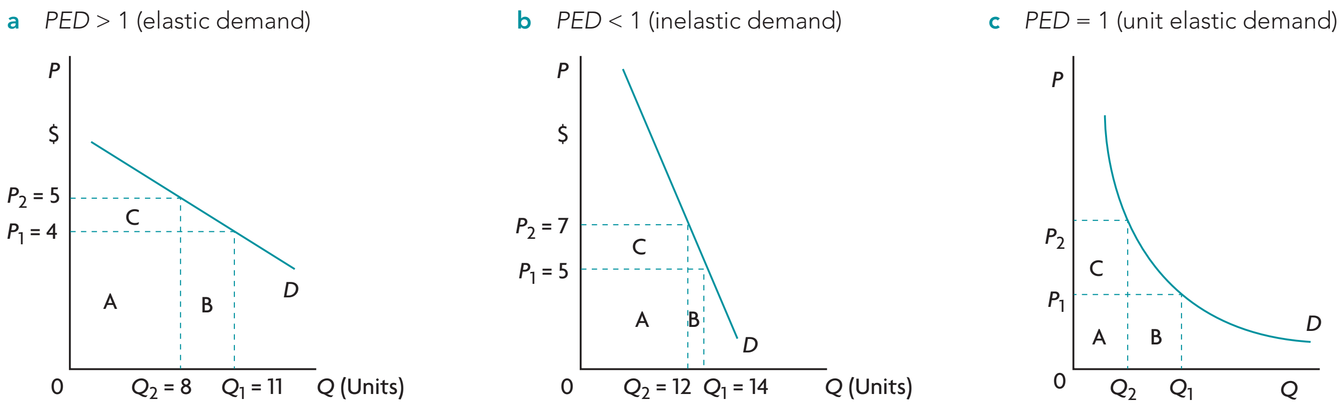


Figure 3.6: PED and total revenue

Calculating total revenue at different points on the demand curve

We can use the information in Figure 3.6 to calculate total revenue as price changes. In Figure 3.6(a) where demand is elastic, initially $TR = P_1 \times Q_1 = 4 \times 11 = \$44$. After the price increase $TR = P_2 \times Q_2 = 5 \times 8 = \$40$. Therefore the increase in price caused a fall in total revenue, confirming the point that when demand is elastic P and TR change in opposite directions.

In Figure 3.6(b) demand is inelastic. Initially $TR = P_1 \times Q_1 = 5 \times 14 = \$70$. After the price increase $TR = P_2 \times Q_2 = 7 \times 12 = \$84$. The price increase here led to an increase in total revenue confirming the point that when demand is inelastic price and total revenue change in the same direction.

PED and firm pricing decisions

The above discussion shows that businesses must take *PED* into account when considering changes in the price of their product. If a business wants to increase total revenue, it must drop its price if demand is elastic, or increase its price if demand is inelastic. If demand is unit elastic, the firm is unable to change its total revenue by changing its price.

A firm's total revenue should not be confused with *profit*. Profit is total revenue minus total costs.

PED and indirect taxes

Governments often impose taxes on specific goods. Such taxes are a type of *indirect tax* (to be discussed in Chapter 4). If governments are interested in increasing their tax revenues, they must consider the *PED* of the goods to be taxed for the following reason.

The lower the price elasticity of demand for the taxed good, the greater the government tax revenues.

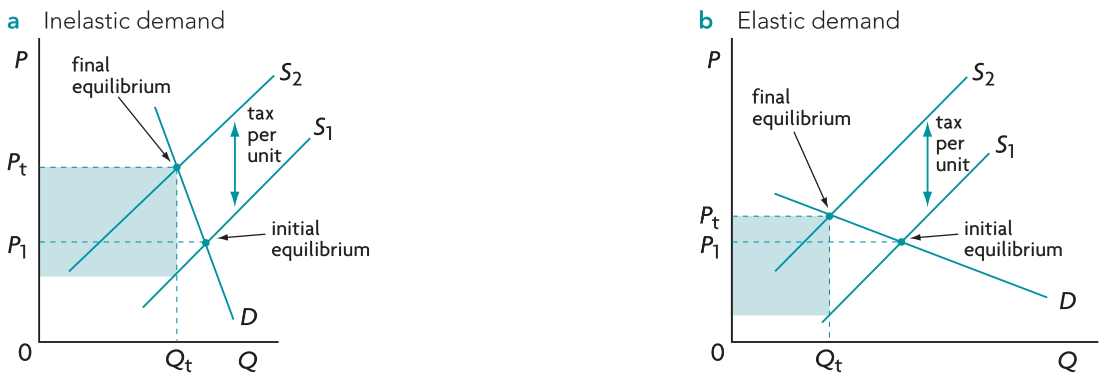


Figure 3.7: PED, indirect taxes and government tax revenue

This can be seen in Figure 3.7, showing the case of inelastic demand in part (a) and elastic demand in part (b).[3] When a tax is imposed on a good, it has the effect of shifting the supply curve upward. The reason is that for every level of output the firm is willing and able to supply to the market, it must receive a price that is higher than the original price by the amount of the tax. (This is equivalent to a leftward shift of the supply curve; for an explanation see 'Quantitative techniques' chapter in the digital coursebook.) The curve shifts from S_1 to S_2 so that the vertical distance between S_1 and S_2 is equal to the amount of the tax per unit of output. The new, after-tax equilibrium occurs at price P_t and quantity Q_t, determined by the intersection of the demand curve, D, and the new supply curve, S_2. The shaded area represents the government's tax revenue, obtained by multiplying the amount of tax per unit times the number of units, or quantity Q_t. A comparison of the two figures indicates that tax revenue is larger when demand is inelastic. This result follows from the principle that when demand is inelastic ($0 < PED < 1$), an increase in price (here due to the increase in the tax) leads to a proportionately smaller decrease in quantity demanded, and hence to an increase in total revenue (i.e. tax revenue). Indirect taxes are therefore usually imposed on goods like cigarettes and petrol (gasoline), which have a low *PED*.

TEST YOUR UNDERSTANDING 3.5

1 Explain and show, using diagrams, how total revenue will change if

- a price increases and demand is elastic,
- b price decreases and demand is inelastic,
- c price increases and demand is perfectly inelastic,
- d price increases and demand is inelastic,
- e price decreases and demand has unit elasticity, and
- f price decreases and demand is elastic.

2 Using diagrams, discuss how a firm's knowledge of price elasticity of demand for its product can help it in its pricing decisions.

3 Refer to Figure 3.2 in this chapter.

- a Calculate the change in total revenue when price increases from \$4 to \$5.
- b Calculate the change in total revenue when price increases from \$1 to \$2.
- c Given your results in (a) and (b), comment on the size of the *PED* for a price change from \$4 to \$5, and the size of the *PED* for a price change from \$1 to \$2.
- d Calculate *PED*s for the two price changes to confirm your answers.

4 (HL only) Referring to Figure 3.4 calculate the change in total revenue that results when:

- a price increases from \$10 to \$15, and
- b price increases from \$40 to \$45.
- c Noting when total revenue increased and when it decreased, explain what your calculations show about *PED* along a straight-line demand curve.

5 The government would like to levy indirect taxes on certain goods to raise tax revenue. Using diagrams, explain how price elasticity of demand can help it decide which products it should tax.

[3] Note we are assuming that the two demand curves are drawn on the same scale, and that if they were drawn in the same diagram they would intersect, therefore the *PED*s are comparable.

PED in relation to primary commodities and manufactured products (HL only)

Why many primary commodities have a lower *PED* compared with the *PED* of manufactured products

Primary commodities are goods arising directly from the use of natural resources, or the factor of production 'land' (see Chapter 1). Primary commodities therefore include agricultural, fishing and forestry products, as well as products of extractive industries (oil, coal, minerals, and so on). Agricultural products include food, as well as other, non-edible commodities (such as cotton and rubber).

Many primary commodities have a low *PED*, which is usually lower than the *PED* of manufactured products (as well as services). **Manufactured products** are goods produced by labour usually working together with capital as well as raw materials, such as for example cars, computers and televisions. Food has a highly price inelastic demand, because it is a necessity and it has no substitutes. The same applies to a variety of other primary products (such as oil and minerals). In the case of food, in developed countries the *PED* is estimated to be between 0.20 and 0.25. By contrast, the demand for manufactured products tends to be more price elastic, because these products, though they may be necessities (in some cases), usually do have substitutes. Therefore, given a price change, quantity demanded is generally more responsive in the case of manufactured products compared with primary commodities. (Note, however, that there are exceptions. For example, medications are manufactured products, yet their demand tends to be inelastic because they are necessities and have no substitutes.)

> Many primary commodities have a relatively low *PED* (price inelastic demand) because they are necessities and have no substitutes (for example, food and oil). The *PED* of manufactured products is relatively high (price elastic demand) because they usually have substitutes.

Consequences of a low *PED* for primary commodities (Supplementary material)

Low price elasticity of demand, together with fluctuations in supply over short periods of time, creates serious problems for primary commodity producers, because they result in large fluctuations in primary commodity prices, and these also affect producers' incomes.

Consider the diagrams in Figure 3.8. Part (a) shows relatively inelastic demand (such as for primary commodities) and part (b) shows relatively elastic demand (such as for manufactured products).

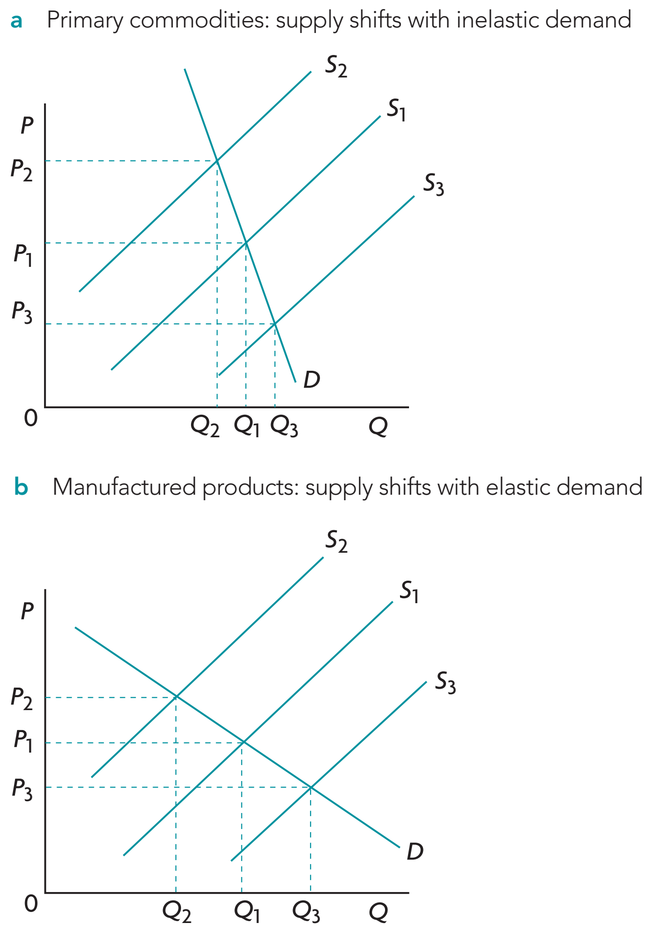


Figure 3.8: Price fluctuations are larger for primary commodities because of low PED

Both diagrams show the effects on price and quantity when there is a decrease in supply (from S_1 to S_2) and when there is an increase in supply (from S_1 to S_3). A comparison of the two diagrams reveals that shifts in the supply curve result in large price fluctuations when demand is inelastic, and much milder ones when demand is elastic. Large price fluctuations over short periods of time are referred to as *price volatility*. (Volatility means instability or high variability.)

Figure 3.8 explains why, in the real world, prices of primary commodities can be highly volatile, whereas they tend to be much less so in the manufacturing and services sectors.

Two results follow from this:

- As primary commodity prices fluctuate widely, so do producers' incomes, which depend on the revenues (price × quantity) producers receive from selling their output.
- In view of the relationship between *PED* and total revenue (see above), a fall in the supply of a primary commodity with inelastic demand (from S_1 to S_2 in Figure 3.8(a)) leads to an increase in total revenue of producers because the percentage increase in price is larger than the percentage decrease in quantity. An increase in supply leads to lower revenues (the percentage decrease in price is larger than the percentage increase in quantity).

These points lead to some unexpected conclusions. They show that a poor crop in agriculture, say due to poor weather conditions, which results in a fall in supply (S_2 in Figure 3.8(a)), leads to higher prices and higher total revenue for farmers. A good crop resulting in a supply increase, or S_3, leads to lower prices and lower farmers' revenues. We come, therefore, to the ironic conclusion that a poor crop may be good for farmers while a good crop may be bad for them.

If supply of agricultural products were relatively stable, the problem would be less serious as agricultural product prices would also be more stable. However, agricultural production depends on many factors beyond the farmer's control, such as drought, pests, floods, frost and other such natural disasters, as well as exceptionally good weather conditions, which occur over short periods of time. These cause frequent and large supply changes (supply curve shifts).

The problem of unstable farmer revenues is an important reason behind government intervention to support farmer incomes, which we will study in Chapter 4. The implications of unstable primary product prices for farmer revenues and the economy will be explored in Chapter 19 .

TEST YOUR UNDERSTANDING 3.6

1 Suppose flooding destroys a substantial portion of this season's crop. Using diagrams, explain what is likely to happen to farmers' revenues, assuming the demand for the product they produce is inelastic.

2 a Using examples, explain why many primary commodities have a relatively low *PED* while many manufactured products have a relatively high *PED*.

b Use the concept of *PED* and diagrams to explain why agricultural product prices tend to fluctuate more (are more volatile) compared with manufactured product prices over the short term. (Optional)

3.2 Income elasticity of demand (*YED*)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- use the formula for income elasticity of demand (*YED*) to calculate *YED*, change in income and quantity demanded (AO4)
- draw an Engel curve diagram to show income elastic, income inelastic and inferior goods (AO4)
- depending on the sign of *YED*, distinguish between normal and inferior goods (AO2)
- depending on the value of *YED* (less than one or greater than one), distinguish between necessities, services and luxury goods (AO2)
- apply *YED* to (HL only) (AO3)
  - discuss its relevance to firms
  - discuss changes in the sectoral structure of an economy

Income elasticity of demand

Understanding income elasticity of demand

Consumer income is an important factor influencing demand for a good and the position of the demand curve. We have already encountered the role of income in Chapter 2, where we saw that income is a factor that causes demand curve shifts (see Chapter 2).

Income elasticity of demand (*YED*) is a measure of the responsiveness of demand to changes in income, and involves demand curve shifts. It provides information on the direction of change of demand given a change in income (increase or decrease) and the size of the change (size of demand curve shifts).

Calculating *YED*, change in price and change in quantity

The formula for *YED* has the same basic form as the other elasticity formulae, and shows the relationship between the percentage change in quantity demanded of a good, *X*, and the percentage change in income, which we abbreviate as *Y*:

income elasticity of demand (*YED*)

$$= \frac{\text{percentage change in quantity demanded of good} \times}{\text{percentage change in income}}$$

$$YED = \frac{\%\Delta Q}{\%\Delta P}$$

Which can be rewritten as:

$$\frac{\frac{\Delta Q}{Q} \times 100}{\frac{\Delta Y}{Y} \times 100} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta Y}{Y}} \text{ where}$$

$\Delta Q = Q_{\text{final value}} - Q_{\text{initial value}}$

$Q = Q_{\text{initial value}}$

$\Delta Y = Y_{\text{final value}} - Y_{\text{initial value}}$

$P = Y_{\text{initial value}}$

Suppose your income increases from \$800 per month to \$1000 per month, and your purchases of clothes increase from \$100 to \$140 per month. What is your income elasticity of demand for clothes?

$$YED = \frac{\frac{40}{100}}{\frac{200}{800}} = \frac{0.40}{0.25} = +1.6$$

Your income elasticity demand for clothes is +1.6.

If we know *YED* and the percentage change in quantity demanded, it is simple to calculate the percentage change in income. Suppose *YED* = 0.75 and quantity demanded of a *normal* good has increased by 15%. The percentage change in income is found by taking

$$YED = 0.75 = \frac{0.15}{\%\Delta Y} \Rightarrow 0.75\text{x}\%\Delta Y = 0.15 \Rightarrow \%\Delta Y$$

= 0.20 or 20%.

Income increased by 20%.

Similarly if we know *YED* and the percentage change in income, we can calculate the percentage change in quantity demanded. Suppose *YED* = 1.25 and income increases by 20%. The percentage change in Q is found by:

$$YED = 1.25 = \frac{\%\Delta Q}{0.20} \Rightarrow \%\Delta Q = 1.25\text{x}0.20 = 0.25 \text{ or } 25\%$$

Quantity demanded increased by 25%.

Interpreting income elasticity of demand

Income elasticity of demand provides two kinds of information:

- the sign of *YED*: positive or negative
- the numerical value of *YED*: whether it is greater or smaller than one (assuming it is positive).

The sign of income elasticity of demand: normal or inferior goods

The sign of *YED* tells us whether a good is normal or inferior:

- *YED* > 0. A positive income elasticity of demand indicates that the good in question is *normal*; demand for the good and income change in the same direction (both increase or both decrease). Most goods are normal goods (see Chapter 2).
- *YED* < 0. A negative income elasticity of demand indicates that the good is *inferior*: demand for the good and income move in opposite directions (as one increases the other decreases). Examples include bus rides, used clothes and used cars; in these cases, as income increases, the demand for these goods fall as consumers switch to consumption of normal goods (new cars, new clothes and so on; see Chapter 2).

The difference between normal and inferior goods can be seen in Figure 3.9, showing a demand curve, D_1, and shifts of the curve that occur in response to increases in income. As income increases, the demand curve shifts rightward from D_1 to D_3 or D_4 when goods are normal ($YED > 0$), but shifts leftward to D_2 when goods are inferior good ($YED < 0$).

The numerical value of income elasticity of demand: necessities, luxuries and services

Here we are making a distinction between goods that have a *YED* that is positive and less than or greater than one:

- *YED* < 1: *Necessities*. If a good has a *YED* that is positive but less than one, it has **income inelastic demand**: a percentage increase in income produces a smaller percentage increase in quantity demanded. Necessities are income inelastic goods.
- *YED* > 1: *Luxuries* and *services*. If a good has a *YED* that is greater than one, it has **income elastic demand**: a percentage increase in income produces a larger percentage increase in quantity demanded. Luxuries and services are income elastic.

Necessities, such as food, clothing and housing, tend to have a *YED* that is positive but less than one; they are normal goods that are income inelastic. In the case of food, as income increases, people buy more food but the amount of income spent on food increases more slowly than income. In developed countries, *YED* for food is about 0.15 to 0.2. This means that a 1% increase in income produces a 0.15% to 0.2% increase in spending on food; or a 10% increase in income results in a 1.5% to 2% increase in spending on food. By contrast, luxuries, such as jewellery and expensive cars, as well as many services, such as travel to other countries, private education and eating in restaurants are income elastic: as income increases, the amount of income spent on such goods increases faster than income (the denominator in the *YED* formula is smaller than the numerator).

What is a necessity and what is a luxury depends on income levels. For people with extremely low incomes, even food and certainly clothing can be luxuries. As income increases, certain items that used to be luxuries become necessities. For example, items like Coca-Cola® and coffee for many poor people in less developed countries are luxuries, whereas for consumers in developed countries they have become necessities. Income elasticity of demand for particular items therefore varies widely depending on income levels. While *YED* for food is about 0.15–0.20 in more developed countries, it is about 0.8 in poor countries. For an increase in income of 10%, spending on food increases by only 1.5%–2% in rich countries and by 8% in poor countries.

In Figure 3.9, we see that in the case of necessities, an increase in income will produce a relatively small rightward shift in the demand curve; in the case of luxuries and services, the rightward shift will be larger.

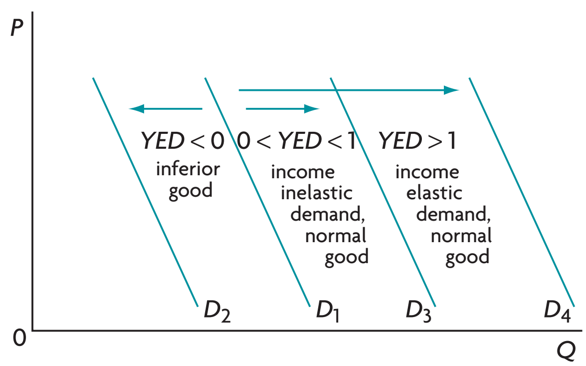


Figure 3.9: Demand curve shifts in response to increases in income for different *YED*s

TEST YOUR UNDERSTANDING 3.7

1. Explain the meaning of income elasticity of demand. Why do we say it involves *a shifting demand curve*?
2. Explain the difference between normal and inferior goods and provide examples of each.
3. Your income increases from £1000 a month to £1200 a month. As a result, you increase your purchases of pizzas from 8 to 12 per month, and you decrease your purchases of cheese sandwiches from 15 to 10 per month.
   - a Calculate your income elasticity of demand for pizzas and for cheese sandwiches.
   - b What kind of goods are pizzas and cheese sandwiches for you?
   - c Show using diagrams the effects of your increase in income on your demand for pizzas and cheese sandwiches.
4. A 15% increase in income leads to a 10% increase in demand for good A and 20% increase in demand for good B.
   - a Identify which of the two goods is income elastic and which is income inelastic.
   - b Identify which of the two goods is likely to be a necessity good and which a luxury good.

The Engel curve

The Engel curve is a far more accurate way to illustrate *YED* than demand curve shifts. It is named after Ernst Engel, a German statistician and economist who lived in the 19th century, and who was the first to study the relationship between consumer income and demand for a product. An Engel curve is shown in Figure 3.10, where the vertical axis measures the income of a consumer per week, the horizontal axis measures the quantity of hot dogs she or he buys each week, while the solid line shows the curve itself. We can see straight away that hot dogs are a normal good from point A to point C, since as income increases from $100 to $250, the quantity increases from 4 hot dogs to 9 hot dogs. As income increases further to $350 the quantity remains constant at 9 hot dogs, but as income increases even more the quantity falls from 9 to 8 hot dogs. We can see therefore that when income goes above $350, hot dogs become an inferior good for this consumer.

With income on the vertical axis and quantity on the horizontal axis of an **Engel curve** diagram, we can see the following:

- YED>0 in the upward sloping part of the curve showing quantity and income both increasing, which indicates the good is normal
- YED<0 in the downward sloping part showing quantity decreasing as income increases, which indicates the good is inferior.

The Engel curve can also show whether a good is income inelastic or income elastic. In Figure 3.10, for very low incomes less than $150, hot dogs are a luxury as *YED*>1. At the higher income levels between $250 and $350, hot dogs have become a necessity as *YED*<1.

There is a simple rule that allows us to distinguish between a luxury and a necessity on the upward sloping part of the Engel curve.

Imagine each segment of the Engel curve extending backward to touch either the vertical axis or the horizontal axis, as shown by the dotted lines:

- *YED*>1 if the line touches the vertical axis, as with the line AB, so that it is a luxury or service
- *YED*<1 if the line touches the horizontal axis, as with line BC, so that it is a necessity.

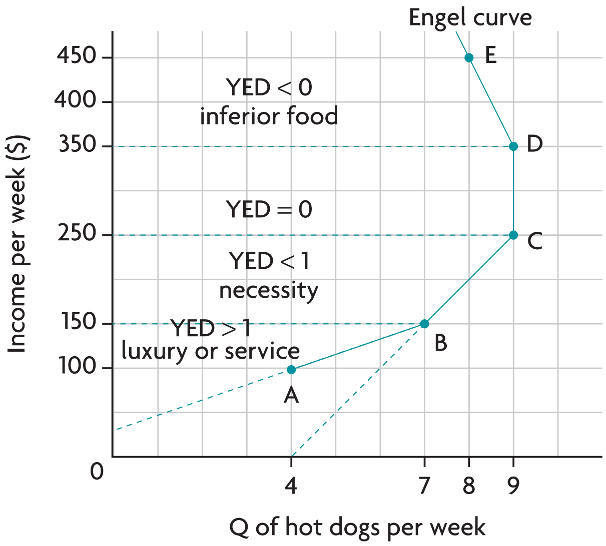


Figure 3.10: The Engel curve showing different YEDs

(The mathematically inclined student can easily see why this is so. As we know

$$YED = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta Q}{Y}} = \frac{\Delta Q}{\Delta Y} \times \frac{Y}{Q} = \frac{\frac{Y}{Q}}{\frac{\Delta Y}{\Delta Q}}$$

Imagine now a line joining the origin with point B on the Engel curve. The slope of this line is Y/Q. Note also that the slope of the Engel curve is ΔY/ ΔQ. Therefore *YED* has been written as the ratio of the two slopes: slope of OB/slope of Engel curve. Now note that OB is steeper than AB, therefore the slope of OB>the slope of the Engel curve from A to B, and so *YED*>1 up to point B. On the other hand, OB is flatter than BC, therefore the slope of OB<the slope of the Engel curve from B to C, and so *YED*<1 from B to C.)

The Engel curve shows a continuum: at very low incomes a good may be a luxury; as income increases it becomes a necessity and finally at high income levels it becomes inferior.

The information in Figure 3.10 can be used to calculate *YED*s in order to confirm the above points. This will be given to you as an exercise.

TEST YOUR UNDERSTANDING 3.8

1 Draw an Engel curve for a typical good and explain how it changes from income elastic to income inelastic to inferior as income increases.

2 Describe why income elasticity of demand for food has been estimated to be about 0.15 to 0.2 in more developed countries and about 0.8 in less developed countries.

3 Using the information in Figure 3.10, calculate *YED* for an increase in income

- a from $100 to $150,
- b $150 to $250,
- c $250 to $350, and
- d $350 to $450.
- e Explain what your results for these *YED*s tell you about the nature of the good for the various income levels.

Applications of income elasticity of demand (HL only)

YED and producers: the rate of expansion of industries

Over time, as countries experience economic growth, society's income increases. Increasing income means a growing demand for goods and services. Suppose that total income in an economy grows at an average rate of about 3% per year. If goods and services have income elastic demand ($YED > 1$), this means that demand for these goods and services grows at a higher rate than 3%. Examples usually include restaurants, movies, health care and foreign travel. Other goods and services have income inelastic demand ($YED < 1$), meaning that the demand for these grows at a rate of less than 3%. Examples include food, clothing and furniture. The first group (with the elastic demand) includes goods and services produced by industries that grow and expand faster than total income in the economy, while the second group includes goods or services produced by industries growing more slowly than total income.

The higher the *YED* for a good or service, the greater the expansion of its market is likely to be in the future;

the lower the *YED*, the smaller the expansion. Producers interested in producing in an expanding market may therefore want to know *YED*s of various goods and services.

In contrast to periods of economic growth, if an economy is experiencing a recession (falling output and incomes, see Chapter 8), goods and services with high *YED*s ($YED>1$) are the hardest hit, experiencing the largest declines in sales. Products with low *YED*s ($YED<1$) can avoid large falls in sales, while inferior goods ($YED<0$) can even experience increases in sales.

YED and the sectoral structure of the economy

The implications of differing *YED*s for the economy follow from what happens to particular industries in the economy as income grows, discussed above.

Every economy has three sectors (or parts): the primary sector including primary products (agriculture, forestry, fishing and extractive industries), the manufacturing sector and the services sector (including entertainment, travel, banking, insurance, health care, education, and so on). With economic growth, the relative size of the three sectors usually changes over time, and these changes can be explained in terms of income elasticity of demand.

Agriculture, the main part of the primary sector, produces food, which as noted above has a *YED* that is positive but less than one (it is income inelastic). As society's income grows over time, the demand for agricultural output grows more slowly than the growth in income. Other primary products also have a low income elasticity of demand. For example, cotton and rubber have synthetic substitutes, so as income increases a relatively larger proportion of it is spent on the synthetic materials, while a relatively lower fraction goes towards cotton and rubber. By contrast, manufactured products (cars, televisions, computers, and so on) have a *YED* that is usually greater than one (income elastic), so that as society's income grows, the demand for these products grows faster than income. Many services have even higher *YED*s, so the percentage increase in the demand for these is much larger.

Therefore, over time, the share of agricultural output in total output in the economy shrinks, while the share of manufactured output grows. With continued growth, the services sector expands at the expense of both agriculture and manufacturing.

REAL WORLD FOCUS 3.2

Income elasticity of demand and growth in demand for food products

According to the World Bank (an international organisation that lends to developing countries, see Chapter 20), the global rate of growth in demand for agricultural commodities like rice and wheat will fall from an average of 2.8% per year in the period 2010–2016 to an average of 1.8% in the period 2017–2018. The reason for this expected trend is that as consumers' incomes increase, consumers switch from commodities like rice and wheat to other foods that have a high protein and fat content.

Applying your skills

1. Explain the meaning of income elastic demand and income inelastic demand.
2. Based on the information in the text, outline your conclusion about the *YED* of commodities like rice and wheat, in comparison with the *YED* of foods with a high protein and fat content. In other words, which of the two is likely to be greater?
3. Explain why goods and services with highly income elastic demand stand to gain the most from rising consumer incomes.

Source: *Manas Chakravarty 'World Bank says long term growth in commodity consumption to weaken', 13 June 2018, Livemint www.livemint.com*

Figure 3.11: A Nepali woman winnowing rice, which is the major crop in Nepal

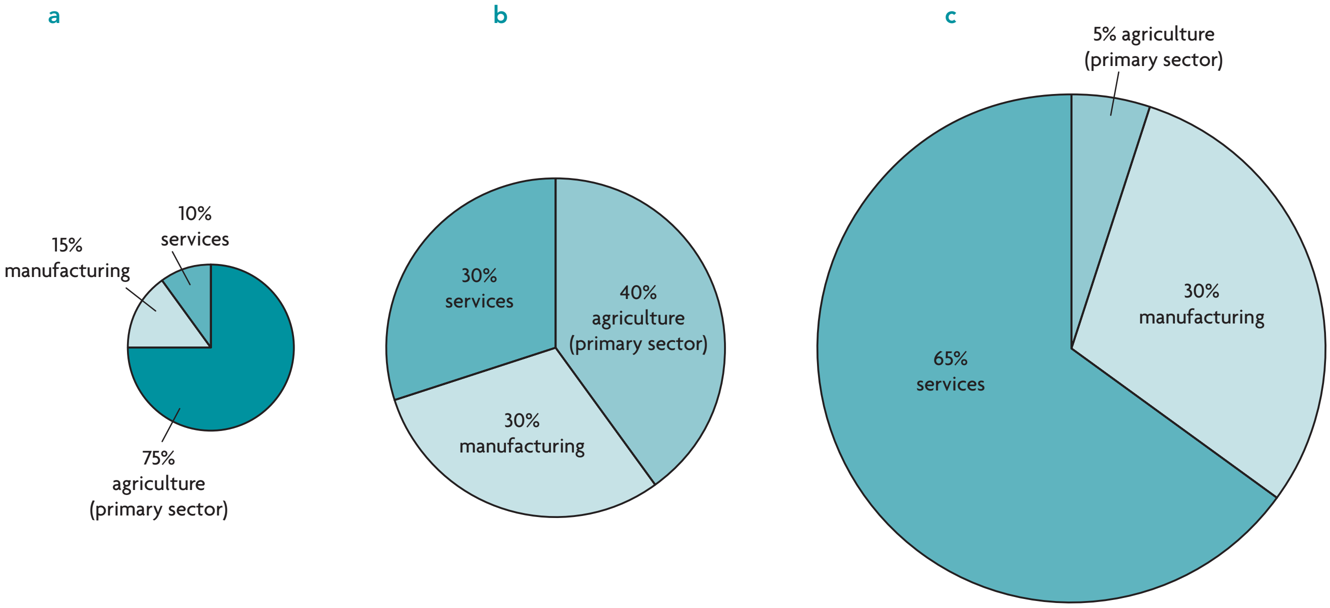


Figure 3.12: Changing relative shares (as percentage of total output) of primary, manufacturing and services sectors for a hypothetical economy as it grows

In Figure 3.12, this is shown for a hypothetical growing economy in the changes of relative sector sizes from parts (a) to (b) to (c).

Economically less developed countries usually have a large primary sector due to the importance of agriculture and extractive activities, while manufacturing and services are far less important. The developed countries of today were in a similar position many decades ago. The historical experience of both more and less developed countries shows that with economic growth, the primary sector becomes less and less important, and is partly replaced by manufacturing and services. As the economy grows further, the relative importance of the primary sector continues to shrink, and manufacturing becomes increasingly replaced by services. Thus, while less developed countries are usually dominated by the primary sector, more developed countries are dominated by services. In the developed world today, among the industries experiencing the fastest growth are services, including education, health care, travel and financial services.

Note that if total output is increasing over time, a falling share for a particular sector (such as the primary sector) does not necessarily mean that primary sector output is falling. Most likely it means that this sector's output is growing, but more slowly than total output. An increasing share for a sector means that its output is growing more rapidly than total output.

TEST YOUR UNDERSTANDING 3.9

1 Discuss why firms would be interested in knowing the *YED* of various goods and services.

2 Explain a likely reason behind the observed rapid growth in certain service industries, including health care, education and financial services, compared with other industries such as food (in the primary sector) and furniture (in the secondary sector).

3 Discuss the role of *YED* in the observed pattern of change in the relative shares of the primary, secondary and tertiary sectors in the economy as a country grows and develops.

3.3 Price elasticity of supply (*PES*)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- use the formula for price elasticity of supply (*PES*) to calculate *PES*, changes in price and changes in quantity (AO4)
- identify the various degrees and range of values for *PES* (AO2)
- draw diagrams showing the range of values for *PES*, including relatively elastic and inelastic supply; and constant values for perfectly elastic supply, perfectly inelastic supply and unitary *PES* (AO4)
- analyse the determinants of *PES* (AO2)
- apply *PES* to analyse the reasons why primary commodities generally have a lower *PES* than manufactured products (HL only) (AO2)

Price elasticity of supply

Understanding price elasticity of supply

Until now, we have studied two demand elasticities, both of which involve consumer responses. We now turn to examine price elasticity of supply, which concerns firm (business) responses to changes in price. According to the law of supply, there is a positive relationship between price and quantity supplied: when price increases, quantity supplied increases and vice versa, *ceteris paribus*. But by how much does quantity supplied change?

Price elasticity of supply (*PES*) is a measure of the responsiveness of the quantity of a good supplied to changes in its price. *PES* is calculated along a given supply curve. In general, if there is a relatively large responsiveness of quantity supplied, supply is referred to as being *elastic*; if there is a relatively small responsiveness, supply is *inelastic*.

Calculating *PES*

The formula for price elasticity of supply (*PES*) follows the same general form of elasticity formulae, only now we consider the relationship between the percentage change in the price of a good, *X*, and the percentage change in quantity of *X* supplied:

price elasticity of supply (*PES*)

$$= \frac{\text{percentage change in quantity of good } X \text{ supplied}}{\text{percentage change in price of good } X}$$

$$PES = \frac{\%\Delta Q}{\%\Delta P}$$

which can be rewritten as

$$\frac{\frac{\Delta Q}{Q} \times 100}{\frac{\Delta P}{P} \times 100} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} \text{ where}$$

$$\Delta Q = Q_{\text{final value}} - Q_{\text{initial value}}$$

$$Q = Q_{\text{initial value}}$$

$$\Delta P = P_{\text{final value}} - P_{\text{initial value}}$$

$$P = P_{\text{initial value}}$$

Calculating PES, change in price and change in quantity

Suppose the price of strawberries increases from €3 per kg to €3.50 per kg, and the quantity of strawberries supplied increases from 1000 to 1100 tonnes per season. Calculate *PES* for strawberries.

$$PES = \frac{\frac{100}{1000}}{\frac{0.50}{3.00}} = \frac{0.10}{0.17} = +0.59$$

Price elasticity of supply for strawberries is +0.59.

If you are given *PES* and the percentage change in quantity supplied, you can find the percentage change in price by solving for this based on the formula above, in exactly the same way as in the case of *PED*. Similarly, you can solve for the percentage change in quantity supplied if you are given *PES* and percentage change in price.

Interpreting price elasticity of supply

The range of values for *PES*

Price elasticity of supply ranges in value from zero to infinity. Because of the positive relationship between price and quantity supplied, *PES* is positive.

The value of *PES* involves a comparison of the percentage change in quantity supplied (the numerator in the formula for *PES*) with the percentage change in price (the denominator). This comparison yields the following possible values and range of values of *PES*, which are illustrated in Figure 3.13 and summarised in Table 3.2:

- **Supply is price inelastic when *PES* < 1**. The percentage change in quantity supplied is smaller than the percentage change in price, so the value of *PES* is less than one; quantity supplied is relatively unresponsive to changes in price, and supply is *price inelastic* or *inelastic*. Figure 3.13(a) shows an inelastic supply curve (*PES* < 1), where a 10% price increase leads to a 5% increase in quantity supplied. When *PES* < 1, the supply curve extends upward and to the right from the horizontal axis; its end-point cuts the horizontal axis.[4]
- **Supply is price elastic when *PES* > 1**. The percentage change in quantity supplied is larger than the percentage change in price, so the value of the *PES* is greater than one; quantity supplied is relatively responsive to price changes, and supply is *price elastic* or *elastic*. Figure 3.13(b) shows an elastic supply curve (*PES* > 1) where the percentage increase in price (10%) is smaller than the percentage increase in quantity (15%). When *PES* > 1, the supply curve extends upward and to the right from the vertical axis; its end-point cuts the vertical axis.[5]

In addition, there are three special cases of constant PES along the length of the supply curve:

- **Supply is unit elastic when PES = 1.** The percentage change in quantity supplied is equal to the percentage change in price, so *PES* is equal to one; supply is *unit elastic;* there is **unitary *PES***. In Figure 3.13(c), all three supply curves shown

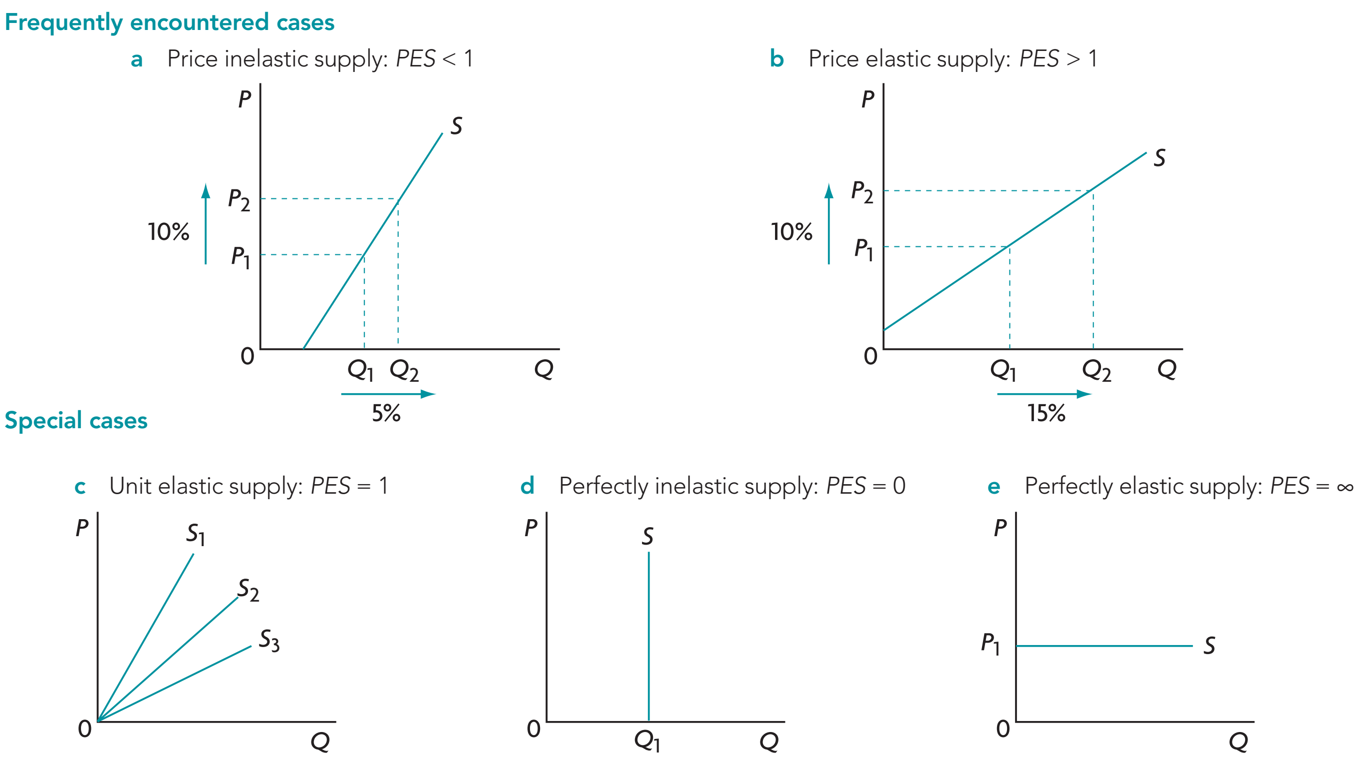


Figure 3.13: Supply curves and *PES*

[4] The reasoning here is exactly the same as in the case of the Engel curve where *YED*<1 if the extension of the Engel curve touches the horizontal axis.

[5] Here, too, the logic is the same as in the Engel curve where *YED*>1 if the Engel curve touches the vertical axis.

are unit elastic supply curves, i.e. for all three, *PES* = 1. Any supply curve that passes through the origin has a *PES* equal to unity. The reason for this is that along any straight line that passes through the origin, between any two points on the line the percentage change in the vertical axis (the price) is equal to the percentage change in the horizontal axis (the quantity). Therefore, for lines that pass through the origin, it is important not to confuse the steepness of the curve with the elasticity of the curve.

- **Supply is perfectly inelastic when PES = 0.** The percentage change in quantity supplied is zero; there is no change in quantity supplied no matter what happens to price; *PES* is equal to zero and supply is said to be **perfectly inelastic**. In Figure 3.13(d), the supply curve is vertical at the point of fixed quantity supplied, Q_1. This is the same as the supply curve shown in Figure 2.7 in Chapter 2. Examples of a vertical supply curve include the supply of fish at the moment when fishing boats return from sea; the season's entire harvest of fresh produce brought to market; the supply of Picasso paintings.
- **Supply is perfectly elastic when PES = ∞.** The percentage change in quantity supplied is infinite; any change in price leads to an infinitely large response in quantity supplied; supply in this case is called **perfectly elastic**, and is shown in Figure 3.13(e) as a horizontal line. (We will encounter such a supply curve in Chapter 14.)

Price elasticities of supply most commonly encountered in the real world are those representing elastic or inelastic supply, with perfectly elastic, perfectly inelastic and unit elastic supply being special cases. Note that only when two supply curves intersect (when they share a price and quantity combination) is it possible to make comparisons of price elasticities of supply by reference to the steepness of the curves. (We have the same condition for making comparisons of *PED*s in the case of demand curves as explained earlier). In the case of intersecting supply curves, the flatter the supply curve, the more elastic it is at any given price. For example, in Figure 3.14, at any one particular price level, S_3 is more elastic than S_2, which is more elastic than S_1.

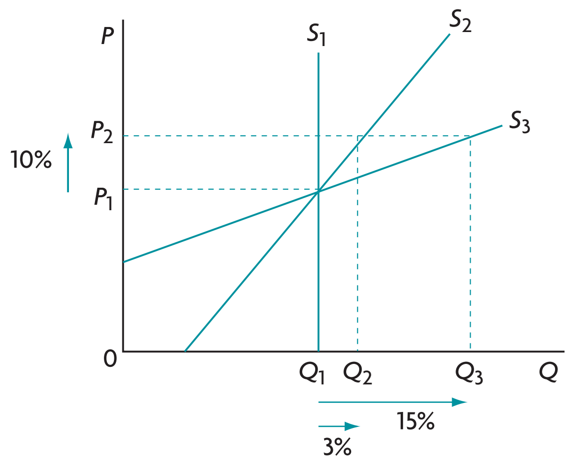


Figure 3.14: The length of time and *PES*

Value of PES	Classification	Interpretation
Frequently encountered cases		
$0 < PES < 1$ (greater than zero and less than one)	inelastic supply	quantity supplied is relatively unresponsive to price
$1 < PES < \infty$ (greater than one and less than infinity)	elastic supply	quantity supplied is relatively responsive to price
Special cases		
$PES = 1$	unit elastic supply	percentage change in quantity supplied equals percentage change in price
$PES = 0$	perfectly inelastic supply	quantity supplied is completely unresponsive to price
$PES = \infty$	perfectly elastic supply	quantity supplied is infinitely responsive to price

Table 3.2: Characteristics of price elasticity of supply

Determinants of price elasticity of supply

We will now consider the factors that determine whether the supply for a good is elastic or inelastic.

Length of time

An important factor determining *PES* is the amount of time firms have to adjust their inputs (resources) and the quantity supplied in response to changes in price. Over a very short time, the firm may be unable to increase or decrease any of its inputs to change the quantity it produces. In this case, supply is highly inelastic, and may even be perfectly inelastic ($PES = 0$). In Figure 3.14, this is represented by S_1. For example, a fishing boat upon its return from a fishing trip has only so many fish to supply in the market. Even if the price of fish rises, there can be no response in quantity supplied. As the length of time that firms have increases, the responsiveness of quantity supplied to price changes begins to rise, and *PES* increases.

In Figure 3.14, the supply curve S_2 corresponds to a time period when the fishing boat can be taken out to sea more often, and more labour can be hired to fish, so as price increases to P_2, quantity supplied increases to Q_2 (the 10% price increase from P_1 to P_2 leads to a 3% increase in quantity supplied, indicating inelastic supply, as $PES < 1$). If an even longer time period goes by, the ability of firms to respond to price changes becomes much greater. The owner of the fishing boat can now not only hire more labour but can also buy more fishing boats, thus greatly increasing the amount of fish that can be supplied. This is shown by the supply curve S_3, for which the price P_2 gives rise to the much larger quantity Q_3 (the 10% price increase from P_1 to P_2 leads to a 15% increase in quantity supplied, indicating elastic supply, as $PES > 1$). Therefore, the larger amount of time firms have to adjust their inputs increases, the larger the *PES*.

Mobility of factors of production

The more easily and quickly resources can be shifted out of one line of production and into another (where price is increasing), the greater the responsiveness of quantity supplied to changes in price, and hence the greater the *PES*. For example, a farm worker can move more easily from strawberry cultivation to corn cultivation than the same farm worker can move to car production.

Spare (unused) capacity of firms

Sometimes firms may have capacity to produce that is not being used (for example, factories or equipment may be idle for some hours each day). If this occurs, it is relatively easy for a firm to respond with increased output to a price rise. But if the firm's capacity is fully used, it will be more difficult to respond to a price rise. The greater the spare (unused) capacity, the higher the *PES* (the more elastic the supply).

Ability to store stocks

Some firms store stocks of output they produce but do not sell right away. Firms that have an ability to store stocks are likely to have a higher *PES* for their products than firms that cannot store stocks.

Rate at which costs increase

If the costs of producing extra output increase rapidly, then supply will be inelastic, as firms will have difficulty expanding their output since they are unlikely to want to incur large costs. On the other hand, if the costs of producing more output rise slowly, it will be easier for firms to expand their output so supply will be elastic. For example, if the price of fertiliser is rising rapidly, thus raising the farm's costs of production, the farmer will find it more difficult to expand output quickly, therefore *PES* is likely to be lower than if the price of fertiliser were stable.

TEST YOUR UNDERSTANDING 3.10

1. a Explain the meaning of price elasticity of supply.
   b Why do we say it measures responsiveness of quantity *along a given supply curve*?
2. Identify the value or range of values for each of the following *PES*s, and show, using diagrams, the shape of the supply curve that corresponds to each one:
   a perfectly elastic supply
   b unit elastic supply
   c perfectly inelastic supply.

CONTINUED

3 a Identify the price elasticity of supply values or range of values that we see most frequently in the real world.

b Compare these by drawing supply curves in a single diagram.

4 Using examples, explain the determinants of *PES*.

5 Suppose that in response to an increase in the price of good *X* from $10 to $15 per unit, the quantity of good *X* produced

a does not respond at all during the first week,

b increases from 10 000 units to 12 000 units over five months, and

c increases from 10 000 to 18 000 units over two years. Calculate *PES* for each of these three time periods and identify when it is price elastic, price inelastic or perfectly inelastic.

6 a Explain what factors can account for the difference in the size of the three elasticities of question 5.

b Draw a supply curve that is likely to correspond to each of the three elasticities in a single diagram.

Applications of price elasticity of supply (HL only)

PES in relation to primary commodities and manufactured products

Why many primary commodities have a lower PES compared with the PES of manufactured products

In general, primary commodities usually have a lower *PES* than manufactured products. The main reason is the time needed for quantity supplied to respond to price changes. In the case of agriculture, it takes a long time for resources to be shifted in and out of agriculture. Farmers need at least a planting season to be able to respond to higher prices. In most areas there is a limited amount of new land that can be brought into cultivation. In some regions of the world land appropriate for agriculture is shrinking due to environmental destruction (caused by over-farming that depletes the soil of minerals needed by crops). Under such conditions, what is needed is an increase in output per unit of land cultivated (crop yields), but this requires technological change in agriculture, involving new seeds or other inputs that are more productive, and takes a great deal of time. Also needed are more and better irrigation systems, although many countries face a growing water shortage. All these factors explain why a long time is needed for the quantity of an agricultural commodity to respond to increases in price.

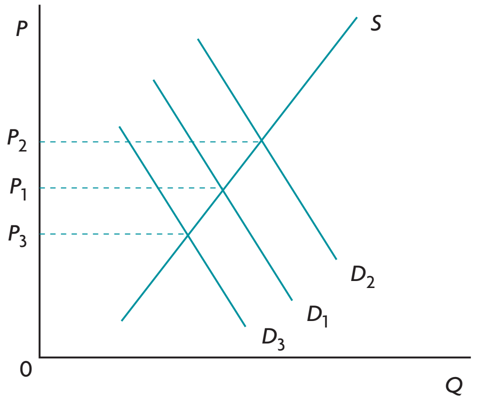
In the case of other primary products, such as oil, natural gas and minerals, time is needed to make the necessary investments and to begin production. Because of the costs involved, firms do not respond quickly to price increases, and wait for a serious shortage (excess demand) in the commodity to arise before they take actions to increase production.

Consequences of a low PES for primary commodities (Supplementary material)

Earlier, in our discussion of price elasticity of demand (*PED*), we saw that price inelastic demand for primary products is an important factor contributing to short-term price and revenue instability for producers such as farmers. Inelastic supply of agricultural and other primary products also contributes to price and income instability for primary product producers.

Figure 3.15 shows a fluctuating demand curve: in part (a) it interacts with inelastic supply, which is typical in the case of primary products, and in part (b) with elastic supply, more typical of manufactured products. Clearly, price fluctuations are larger in the case of inelastic supply. Large price fluctuations mean large revenue fluctuations, or unstable revenue for producers of primary commodities. We will come back to the implications of unstable prices and revenues for producers and for the economy in Chapters 4 and 19.

a Primary commodities: demand shifts with inelastic supply



b Manufactured products: demand shifts with elastic supply

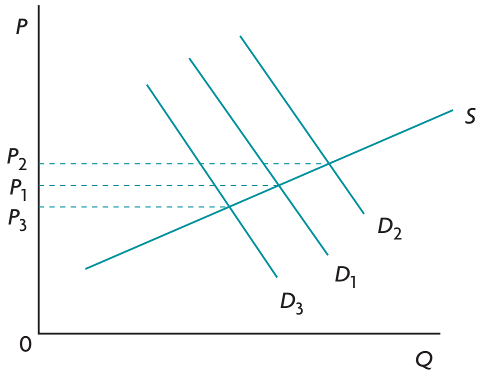


Figure 3.15: Price fluctuations are larger for primary commodities because of low *PES*

Short-term and long-term price elasticities of supply

It was noted above that agricultural products (and other primary commodities) usually have lower price elasticities of supply than manufactured products because they need more time to respond to price changes. This suggests that over longer periods of time the *PES* of agricultural products is larger.

Table 3.3 shows that this is in fact the case. The longer the time producers have to make the necessary adjustments, the greater the responsiveness of quantity supplied to price changes (see Figure 3.14).

Commodity	Short-term *PES*	Long-term *PES*
Cabbage	0.36	1.20
Carrots	0.14	1.00
Cucumbers	0.29	2.20
Onions	0.34	1.00
Green peas	0.31	4.40
Tomatoes	0.16	0.90
Cauliflower	0.14	1.10
Celery	0.14	0.95

Table 3.3: Short-run and long-run *PES* for selected agricultural commodities

TEST YOUR UNDERSTANDING 3.11

1 a Explain why the *PES* for many primary commodities is relatively low and for many manufactured products is relatively higher.

b Use the concept of *PES* to explain why agricultural product prices are volatile over the short term. (Optional)

Summary of *PED*, *YED* and *PES*

Table 3.4 provides a summary of key characteristics of all the elasticities considered in this chapter.

Elasticity	Possible values	Description	Examples
Price elasticity of demand $PED = \frac{\%\Delta Q}{\%\Delta P}$	$PED = 0$	perfectly inelastic	concept used in economic theory
	$0<PED<1$	price inelastic	gasoline, cigarettes, food
	$PED = 1$	unit elastic	concept used in economic theory
	$PED>1$	price elastic	yachts, expensive holidays
	$PED = \infty$	perfectly elastic	concept used in economic theory

Elasticity	Possible values	Description	Examples
Income elasticity of demand $YED = \frac{\%\Delta Q}{\%\Delta Y}$	$YED > 0$	normal good	new cars, new clothes
	$YED < 0$	inferior good	used cars, used clothes
	$YED > 1$	income elastic, luxury	expensive cars and clothes, many services
	$YED < 1$	income inelastic, necessity	food, medicines
Price elasticity of supply $PES = \frac{\%\Delta Q}{\%\Delta P}$	$PES = 0$	perfectly inelastic	concept used in economic theory
	$PES < 1$	price inelastic	oil and gasoline, some agricultural products
	$PES = 1$	unit elastic	concept used in economic theory
	$PES > 1$	price elastic	any good that can be produced quickly
	$PES = \infty$	perfectly elastic	concept used in economic theory

Table 3.4: Elasticity concepts: a summary

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1. Research and find products that are subject to indirect taxes in the country you live in. Do you think they are likely to be price elastic or price inelastic? Is the government likely to have high or low revenues arising from the indirect taxes?
2. Think of some goods that are likely to have changing *YED*s on account of their being luxuries, necessities and inferior goods depending on the income level.
3. Consider a good with a *PES* that ranges from zero, positive but less than one, and greater than one depending on the time period and explain why *PES* varies. Try to consider as many determinants of supply to explain your answer.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Real world issue 2: When do markets fail to meet important economic objectives, and what can government intervention offer?

CONCEPTUAL UNDERSTANDINGS

1. The market is often unable to achieve socially desirable outcomes, resulting in reduced ***efficiency***, environmental ***sustainability*** and ***equity***.
2. The issues arising from the inability of the market to achieve socially desirable outcomes include
   a. market failure, where the market fails to achieve allocative ***efficiency*** thus resulting in welfare loss
   b. overuse of resources, that challenges environmental ***sustainability***
   c. inequity, which results in inequalities.
3. There are numerous policy tools that governments have at their disposal to deal with the above issues; government ***intervention*** works to change market outcomes, and has varying degrees of effectiveness in a variety of real-world situations.

We have seen in previous chapters that markets are a mechanism for allocating resources and distributing (rationing) output among users. But markets sometimes do not give rise to socially desirable results. Governments, therefore, intervene in markets in order to deal with such issues as pollution, destruction of the environment, provision of more schools and health services, prevention of monopolies and greater fairness in how income is distributed.

Deforestation in the Amazon

Chapter 4

Government intervention in microeconomics

BEFORE YOU START

- Sometimes governments tax goods and services. What might be some reasons for doing so?
- Sometimes governments give money to (subsidise) certain producers. Why do you think government would do this?
- Besides taxing and subsidising, can you think of other ways that governments intervene in markets?

This chapter will examine the following types of government intervention in markets:

- price controls:
  - price ceilings
  - price floors
- indirect taxes
- subsidies.

We will see what effects these policies have on markets and we will evaluate their effects on stakeholders.

Photo: Berlin, Germany, January 2019. Farmers protest against distribution of EU subsidies: large farms receive too much

4.1 Government intervention in markets

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- outline key reasons why governments intervene in markets (AO2)

Why governments intervene in markets

In Chapter 1, we saw that an important justification for government intervention is to compensate for the inability of markets to carry out all socially desirable economic activities effectively. This section will provide a review of the main reasons for government intervention at the microeconomic level. Intervention at the macroeconomic level will be studied in Chapter 13.

Earn revenue for the government

Governments earn revenues from *indirect taxes*, which are taxes on goods and services. Note that these kinds of taxes are different from income taxes, which you will learn about in Chapter 12. In Chapter 3 we saw that the lower the price elasticity of demand for a good, the greater the amount of tax revenue earned. This explains why indirect taxes are often imposed on goods that have a price inelastic demand, meaning $0<PED<1$, such as cigarettes, alcohol and petrol/gasoline.

Provide support to firms

Governments provide support to firms for several reasons. Small firms that have just been set up may require financial assistance to become well established so they can compete with larger firms. Support may also be offered to firms in an industry whose growth the government would like to encourage, such as industries that produce environmentally-friendly forms of energy (for example, wind power and solar power). In these two cases, the government may offer *subsidies* or *price floors*, both of which we will study in this chapter. Further, the government may want to protect domestic firms from foreign competition arising from imports, in which case it may offer trade protection measures such as *tariffs* and *quotas*, to be studied in Chapter 14.

Provide support to households on low incomes

Households often do not earn enough income to be able to provide all necessities to meet their basic needs (food, shelter, clothing). There are several methods the government can use to support such households, including *subsidies*, *price ceilings* and *direct provision of services*, such as the provision of free education and free health care, all of which we will study in this chapter. Also, it provides *transfer payments*, which include such payments as unemployment benefits, child benefits, maternity benefits and many more; these will be considered in Chapters 12 and 20.

Influence the levels of production of firms

When the government provides support to firms through any of the methods mentioned above (subsidies, price floors or trade protection measures), one of the consequences is to also to increase the firms' level of production. On the other hand, *indirect taxes* have the opposite effect; they work to decrease firms' level of output.

Influence levels of consumption of households/consumers

In some cases, the government would like to influence consumers to consume greater quantities of goods and services that are held to be beneficial to them (known as *merit goods*) or to reduce consumption of goods and services held to be harmful (*demerit goods*). Examples of merit goods are education and health care; examples of demerit goods are cigarettes and fatty foods. Merit and demerit goods will be explained in Chapters 5 and 6. To increase consumption levels, the government can use *subsidies* and *direct provision of services* (to be discussed in this chapter), or *nudges* (see also Chapter 2), or *command and control methods*, such as compulsory education up to a certain age, also to be discussed in this chapter. To reduce consumption levels, the government can use *indirect taxes* which raise the price of the good or service, thus lowering the quantity demanded, or *nudges*, or *command and control methods* like prohibiting smoking in public places.

Note that **command and control** refers to government laws and regulations that must be followed. It refers to the command approach to decision-making explained in Chapter 1.

Correct market failure

Market failure refers to the failure of the market to achieve allocative efficiency (introduced briefly in Chapter 2 and to be discussed at length in Chapters 5–7). When market failure occurs, it means that the market produces quantities of a good or service that are too large or too small in relation to what society mostly prefers. In addition, it may mean that certain goods that are socially desirable are not produced at all by the market. When market failure occurs, government intervention is required in order to try to correct it, so that the economy will come closer to achieving socially desirable results. Policies the government can use to deal with issues like this include *indirect taxes*, *subsidies*, *nudges, direct provision of services* and *command and control methods*. All these will be considered in Chapters 5–7.

Promote equity (equality)

The market system as a rule does not achieve an equitable (or fair) distribution of income and wealth. As we saw in Chapter 1, economists usually interpret equity to mean equality. Most societies consider that the market system results in income and wealth distributions that are too unequal, meaning that a relatively small proportion of the population receives too large a share of income, and the reverse, where a large proportion of the population receives only a relatively small share of income. Governments therefore undertake to redistribute income. Some policies that we will consider in this chapter include *price ceilings and subsidies*. Other important policies will be considered in Chapters 12 and 20.

What forms does government intervention in markets take at the microeconomic level?

The following forms of government intervention are undertaken at the microeconomic level. When we say that they operate on the microeconomic level, what we mean is that *each of these works to influence demand or supply for a good or service, thus affecting market outcomes*. The main forms of these interventions include the following:

- price controls:
  - price ceilings
  - price floors
- indirect taxes
- subsidies
- direct provision of services
- command and control regulation and legislation
- consumer nudges (explained in Chapter 2 at HL).

This chapter will consider price controls, indirect taxes and subsidies.

4.2 Price controls

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the consequences of price ceilings and price floors on markets and stakeholders (AO2)
- draw diagrams to illustrate the effects of price ceilings and price floors on markets and stakeholders (AO4)
- evaluate the effects of price ceilings and price floors on markets and stakeholders (AO3)
- calculate the effects of price ceilings and price floors on markets and stakeholders (HL only) (AO4)

Introduction to price controls

The first type of intervention we will consider involves price controls.

Price controls refer to the setting of minimum or maximum prices by the government (or private organisations) so that prices are unable to adjust to their equilibrium level determined by demand and supply. Price controls result in market disequilibrium, and therefore in shortages (excess demand) or surpluses (excess supply).

Price controls differ in a fundamental way from other types of government intervention that we will study below (indirect taxes and subsidies). When a tax is imposed or a subsidy granted, the market *settles at a new equilibrium*, where there is a balance of demand with the new supply. Price controls differ because, once they are imposed, they do not allow a new equilibrium to be established, and instead force a situation where there is *persisting market disequilibrium*.

Market disequilibrium means that the market is prevented from reaching a market-clearing price, and there emerge shortages (excess demand) or surpluses (excess supply) (see Chapter 2).

In the discussion that follows, it is important to bear in mind that the term *surplus* has two different meanings. In one sense it refers to excess supply resulting when quantity supplied is greater than quantity demanded, as we discussed in Chapter 2. In the second sense it refers to the benefits that consumers or producers receive from buying or selling (see Chapter 2). This is not as confusing as it may sound, because *surplus* in the second sense is referred to as 'consumer surplus', or 'producer surplus' or 'social surplus' (or 'community surplus').

Price ceilings: setting a legal maximum price

What is a price ceiling?

A government may in some situations set a legal *maximum price* for a particular good; this is called a *price ceiling*. It means that the price that can be legally charged by sellers of the good must not be higher than the legal maximum price. Price ceilings are usually set in order to make certain goods more affordable to people on low incomes.

Figure 4.1 shows how this works. The equilibrium price is P_e, determined by the forces of demand and supply. The price ceiling, P_c, is set by the government at a level below the equilibrium price, leading to a shortage (excess demand), since quantity demanded, Q_d, is greater than quantity supplied, Q_s. If the market were free, the forces of demand and supply would force price up to P_e. However, now this cannot happen, because the price hits the legally set price ceiling.

Note that to have an effect, the price ceiling must be *below* the equilibrium price. If it were higher than the equilibrium price, the market would achieve equilibrium, and the price ceiling would have no effect.

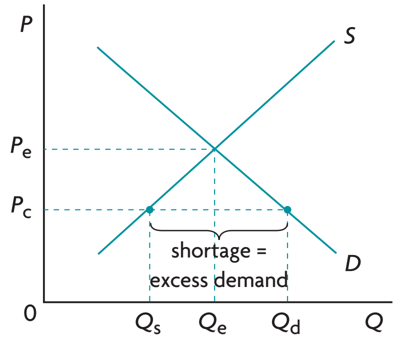


Figure 4.1: Price ceiling (maximum price) and market outcomes

A **price ceiling** is a **maximum price** set *below* the equilibrium price, in order to make goods more affordable to people on low incomes.

Consequences of price ceilings

By imposing a price that is below the equilibrium price, a price ceiling results in a lower quantity supplied and sold than at the equilibrium price. This is shown in Figure 4.1, where the price ceiling, P_c, gives rise to quantity, Q_s, that firms supply, which is less than the equilibrium quantity, Q_e, that suppliers would supply at price P_e.

In addition, the price ceiling, P_c, gives rise to a larger quantity demanded than at the equilibrium price: the quantity consumers want to buy at price P_c is given by Q_d, which is greater than quantity, Q_e, that they would buy at price P_e.

A price ceiling does not allow the market to clear; it creates a situation of disequilibrium where there is a shortage (excess demand).

Consequences for markets

Shortages

A price ceiling, P_c, set below the equilibrium price of a good creates a shortage. At P_c, not all interested buyers who are willing and able to buy the good are able to do so because there is not enough of the good being supplied. In Figure 4.1, the shortage is equal to $Q_d - Q_s$.

Non-price rationing

The term *rationing* refers to a method of dividing up something among possible users (see Chapter 2). In a free market, this is achieved by the price system: those who are willing and able to pay for a good will do so, and the good is rationed among users according to who buys it; this is called *price rationing*. However, once a shortage arises due to a price ceiling, the price mechanism no longer achieves its rationing function. Some demanders willing and able to buy the good at P_c in Figure 4.1 will go unsatisfied. How will the quantity Q_s be distributed among all interested buyers? This can only be done through *non-price rationing* methods, which include:

- waiting in line and the first-come-first-served principle: those who come first will buy the good
- the distribution of coupons to all interested buyers, so that they can purchase a fixed amount of the good in a given time period
- favouritism: the sellers can sell the good to their preferred customers.

Underground (or parallel) markets

Underground (or *parallel*) *markets* involve buying/selling transactions that are unrecorded, and are usually illegal. In the case of price ceilings, they are a special kind of price rationing. They involve buying a good at the maximum legal price, and then illegally reselling it at a price above the legal maximum. Underground markets can arise when there are dissatisfied people who have not succeeded in buying the good because there was not enough of it, and are willing to pay more than the ceiling price to get it. If there were no shortage, the price of the good would be at its equilibrium price, and no one would be interested in paying a higher than equilibrium price for it. Underground markets are inequitable, and frustrate the objective sought by the price ceiling, which is to set a maximum price.

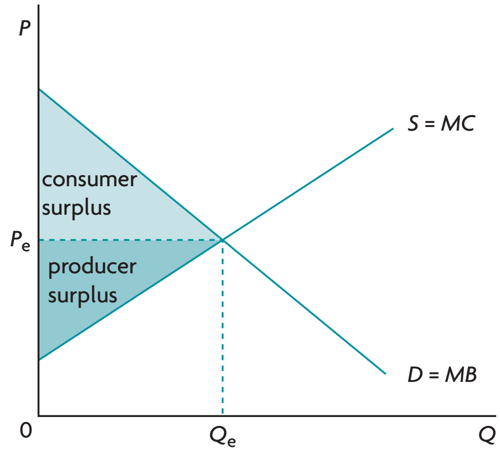
Underallocation of resources to the good and allocative inefficiency

The price ceiling, being lower than the equilibrium price, results in a smaller quantity supplied; in Figure 4.1 $Q_s < Q_e$, not enough resources are allocated to the production of the good, resulting in underproduction relative to the social optimum (or 'best'). Society is worse off due to underallocation of resources and allocative inefficiency.

Negative welfare impacts

Allocative efficiency, and the conditions of maximum social surplus and $MC = MB$ were explained in Chapter 2. In a competitive free market equilibrium, shown in Figure 4.2(a), consumer surplus appears as the shaded area above price P_e and under the demand curve up to quantity Q_e; producer surplus is the shaded area above the supply curve and under price P_e up to Q_e. At the competitive free market equilibrium, the sum of consumer and producer surplus, or social surplus, is maximum, and $MB = MC$, indicating that allocative efficiency is achieved.

a Consumer and producer surplus in a competitive free market: maximum social surplus



b Welfare impacts of a price ceiling (maximum price)

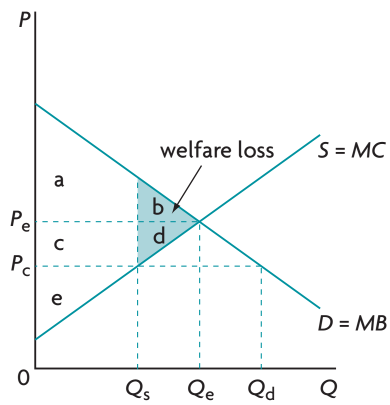


Figure 4.2: Effects of a price ceiling (maximum price) on consumer and producer surplus

In Figure 4.2(b), if there were no price control, the market would determine price P_e and quantity Q_e at equilibrium. Consumer surplus would be equal to areas a + b. Producer surplus , would be equal to areas c + d + e. Consumer plus producer surplus would be maximum, and equal to a + b + c + d + e. Also, $MB = MC$, and there would be allocative efficiency.

If a price ceiling, P_c, is imposed, only the quantity Q_s is produced and consumed. Consumer surplus is now the area under the demand curve and above P_c, but only up to Q_s, since that is all that is consumed. Therefore, consumer surplus becomes a + c. Producer surplus is the area above the supply curve and below P_c, also only up to Q_s since that is all that is produced. Producer surplus therefore falls to area e. Total social surplus after the price ceiling is a + c + e. Comparing with total social surplus before the price ceiling, we see that the shaded areas b and d have been lost and represent *welfare loss*. Welfare loss represents benefits that are lost to society because of resource misallocation.

> **Welfare loss** (also known as *deadweight loss*) represents social surplus or welfare benefits that are lost to society because resources are not allocated efficiently.

We can see there is allocative inefficiency also because MB > MC at the point of production, Q_s: the benefit consumers receive from the last unit of the good they buy is greater than the marginal cost of producing it. Therefore, society is not getting enough of the good, as there is an underallocation of resources to its production.

> A *price ceiling* creates a welfare loss, indicating that the price ceiling introduces allocative inefficiency due to an underallocation of resources to the production of the good. This is shown by $Q_s < Q_e$. Also, $MB > MC$, indicating that society is not getting enough of the good.

Consequences for stakeholders

Stakeholders are individuals or groups of individuals who have an interest in something and are affected by it.

Consumers

Consumers partly gain and partly lose. They gain area c from producers but they lose area b (see Figure 4.2(b)). This is because those consumers who are able to buy the good at the lower price are better off. However, some consumers remain unsatisfied since they don't get to buy the good at all, because at the ceiling price there is not enough of the good to satisfy all demanders.

Producers

Producers are worse off, because with the price ceiling they sell a smaller quantity of the good at a lower price; therefore, their revenues drop from $P_e \times Q_e$ to $P_c \times Q_s$. This is clear also from their loss of some producer surplus, area c (which is transferred to consumers), as well as area d (welfare loss) in Figure 4.2(b).

Workers

The fall in output (from Q_e to Q_s) means that some workers are likely to be fired, resulting in unemployment; clearly these workers will be worse off.

Government

There will be no gains or losses for the government budget, yet the government may gain in political popularity among the consumers who are better off due to the price ceiling.

The examples of rent controls and food price controls

Price ceilings are for the most part set in order to make certain goods considered to be necessities more affordable to low-income earners.

Rent controls

Rent controls consist of a maximum legal rent on housing, which is below the market-determined level of rent (the price of rental housing). It is undertaken by governments in some cities around the world to make housing more affordable to low-income earners. Consequences of rent controls include:

- housing becomes more affordable to low-income earners
- a shortage of housing, as the quantity of housing demanded at the legally maximum rent is greater than the quantity available
- long waiting lists of interested tenants waiting for their turn to secure an apartment/flat
- a market for rented units where tenants sublet their apartments at rents above the legal maximum (an underground market)
- run-down and poorly maintained rental housing because it is unprofitable for landlords to maintain or renovate their rental units since low rents result in low revenues.

Food price controls

Some governments use food price controls as a method to make food more affordable to low-income earners, especially during times when food prices are rising rapidly. The results of food price controls follow the same patterns as discussed above: lower food prices and greater affordability; food shortages as quantity demanded is greater than quantity supplied; non-price rationing methods (such as queues) to deal with the shortages; development of underground markets; falling farmer incomes due to lower revenues; more unemployment in the agricultural sector; misallocation of resources; possible greater popularity for the government among consumers who benefit.

REAL WORLD FOCUS 4.1

Price controls in Vietnam

Figure 4.3: Vietnam. Transport on the Saigon River

Some years ago, due to high rates of inflation (a rising general price level), the Vietnamese government considered the imposition of price controls. These consisted of price ceilings on numerous products, including chemical fertilisers, salt, milk powder, rice, sugar, animal feeds, coal, cement, paper, textbooks and many more. If it introduced these measures, the pricing rules would apply not just to domestic government-owned businesses but also private firms and foreign-owned businesses. It was feared that Vietnam might be moving away from its freer market orientation of recent years and back toward the ways of a command economy. Foreign diplomats warned the government that price controls would damage business confidence in the country.

Applying your skills

1 Outline what it means to move toward the ways of a 'command economy'.

2 Discuss the consequences for the economy and stakeholders that might have arisen if the government imposed the price controls.

3 Why do you think that price controls could damage business confidence?

Source: *Adapted from The Economist Intelligence Unit, 'Vietnam economy: reform roll-back?' in ViewsWire News Analysis, 21 September 2010.*

TEST YOUR UNDERSTANDING 4.1

1 Define a price ceiling and, providing examples, outline some reasons why governments impose them.

2 Using a diagram, explain why price controls lead to disequilibrium market outcomes.

3 Draw a diagram illustrating a price ceiling, and analyse its effects on market outcomes (price, quantity demanded, quantity supplied, market disequilibrium) and consequences for the economy (shortages, non-price rationing, allocative inefficiency, welfare loss).

4 a Explain the difference between price rationing and non-price rationing.

b Describe the circumstances under which non-price rationing arises.

c Identify some forms of non-price rationing.

d Outline why underground markets are a form of price rationing.

5 a Draw a diagram showing producer and consumer surplus in a free market competitive equilibrium.

CONTINUED

b Assuming a price ceiling is imposed in this market, draw a new diagram showing the new consumer surplus, producer surplus and welfare loss.

c Comparing your diagrams for parts (a) and (b), what can you conclude about consumer surplus, producer surplus and welfare loss?

d Describe the relationship between marginal benefits and marginal costs at the new equilibrium. Outline what this reveals about allocative efficiency (or inefficiency).

6 Examine the consequences of price ceilings for different stakeholders in the case of:

a rent controls, and

b food price controls.

Calculating the effects of price ceilings on stakeholders and welfare (HL only)

Figure 4.4 provides us with a numerical example of a price ceiling. At equilibrium, price is equal to £8 and quantity demanded and supplied is 20 000 units of the good per week. When a price ceiling is imposed at P_c = £5.00 per unit, quantity demanded becomes Q_d = 30 000 units, and quantity supplied Q_s = 10 000 units.

Shortage (excess demand)

The shortage, or excess demand, is equal to $Q_d - Q_s$, which in this case is 30 000 – 10 000 = 20 000 units per week.

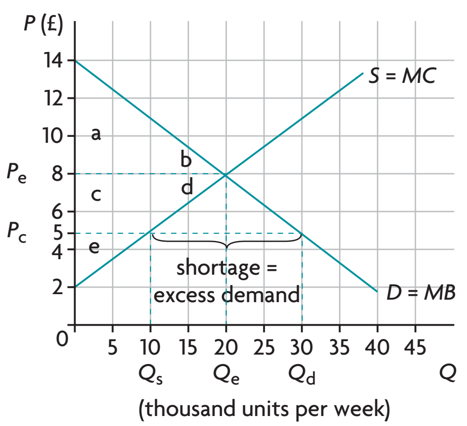


Figure 4.4: Calculating effects of a price ceiling

Change in consumer expenditure

Consumer expenditure is given by the price per unit of the good times the number of units purchased.

At equilibrium, prior to the price ceiling, consumers spend $P_e \times Q_e$ = £8 × 20 000 units = £160 000. After the price ceiling is imposed, consumers spend $P_c \times Q_s$ = £5.00 × 10 000 units = £50 000. The change is therefore £160 000 – £50 000 = £110 000, meaning that consumers now spend £110 000 less than at equilibrium.

Change in producer (firm) revenue

Firm revenue is the same as consumer expenditure both before and after the imposition of the price ceiling. This is because revenue is equal to price per unit times quantity of units sold, and both the price (P_c) and the quantity (Q_s) are the same for both consumers and producers. Therefore, before the price ceiling is imposed, firm revenue is £160 000, and after the price ceiling, firm revenue is reduced to £50 000. Therefore, firm revenues fall by the amount £110 000.

Note that the change in consumer expenditure and the change in firm revenue are the same, because price and quantity were the same for consumers and firms both before and after the price ceiling was imposed.

Change in consumer and producer surplus and welfare loss

In Chapter 2, we saw how consumer and producer surplus are calculated. You may remember that at market equilibrium consumer surplus is half the area of the rectangle whose one side equals the *P* intercept of the demand curve minus the price paid by consumers, and whose other side equals the number of units purchased:

$$\text{Consumer surplus} = \frac{(P \text{ intercept of D curve minus } P \text{ of consumers}) \times Q \text{ purchased}}{2}$$

Therefore, using the information in Figure 4.4, consumer surplus before the price ceiling is:

$$\text{Initial consumer surplus} = \frac{(14-8)\times 20\,000}{2} = \frac{6\times 20\,000}{2}$$

$= £60\,000$

Now notice that after the price ceiling is imposed, consumer surplus no longer has the area of a triangle. It consists of a trapezium composed of areas a + c. The formula for a trapezium was also presented in Chapter 2. *It is the sum of the two parallel sides times the distance between them divided by 2*. Therefore:

$$\text{Final consumer surplus} = \frac{[(14-5)+(11-5)]\times 10\,000}{2}$$

$$= \frac{(9+6)\times 10\,000}{2} = \frac{150\,000}{2} = £75\,000$$

Therefore consumer surplus has increased

by $£75\,000 - £60\,000 = £15\,000$.

Producer surplus at market equilibrium is half the area of the rectangle whose one side equals the price received by producers minus the P intercept of the *initial* supply curve, S_1 and whose other side equals the number of units sold:

Producer surplus =

$$\frac{(P \text{ of producers minus } P \text{ intercept of } S \text{ curve})\times Q \text{ sold}}{2}$$

Therefore producer surplus before the price ceiling is:

$$\text{Initial producer surplus} = \frac{(8-2)\times 20\,000}{2} = \frac{6\times 20\,000}{2}$$

$= £60\,000$

Notice that after the price ceiling is imposed producer surplus is the area of triangle e therefore we can apply the triangle formula:

$$\text{Final producer surplus} = \frac{(5-2)\times 10\,000}{2} = \frac{3\times 10\,000}{2}$$

$= £15\,000$

Producer surplus has decreased by $£60\,000 - £15\,000 = £45\,000$.

The welfare loss is given by the area of triangles a plus b:

$$\text{Welfare loss} = \frac{(11-5)\times(20\,000-10\,000)}{2} = \frac{6\times 10\,000}{2}$$

$= £30\,000$

TEST YOUR UNDERSTANDING 4.2

1 In the example of the market illustrated in Figure 4.4, comment on the effect of a price ceiling set at £10.

2 The diagram below shows a price ceiling of €30 that has been set for good X. Calculate:

a the shortage (excess demand),

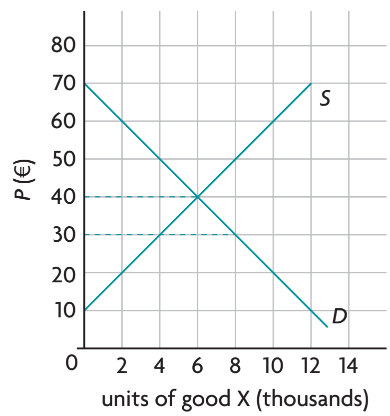
b the change in consumer expenditure,

c the change in producer revenue,

d the change in consumer surplus,

e the change in producer surplus, and

f welfare loss.



Price floors: setting a legal minimum price

What is a price floor?

A legally set *minimum price* is called a *price floor*. The price that can be legally charged by sellers of the good must not be lower than the price floor, or minimum price. In Figure 4.5, a price floor, P_f, is set above the equilibrium price, P_e. At P_e, consumers are willing and able to buy Q_d of the good, but firms are willing and able to supply Q_s of the good. Therefore, a surplus, or excess supply, equal to the difference between Q_s and Q_d, arises. If the market were free, the forces of demand and supply would force the price down to P_e. However, now this cannot happen.

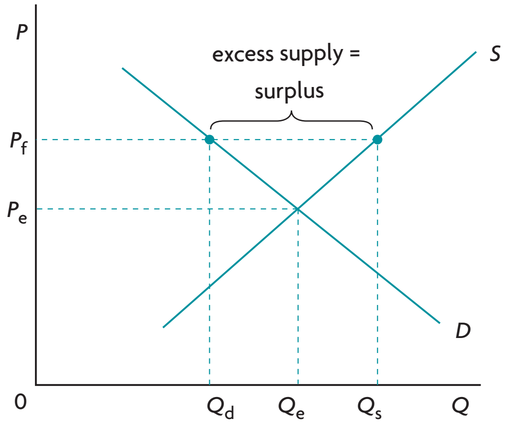


Figure 4.5: Price floor (minimum price) and market outcomes

Note that to have an effect, the price floor must be *above* the equilibrium price. If it were below the equilibrium price, the market would achieve equilibrium and the price floor would have no effect.

Why governments impose price floors

Price floors are commonly used for two reasons: (a) to provide income support for farmers by offering them prices for their products that are above market-determined prices; and (b) to protect low-skilled, low-wage workers by offering them a wage (the minimum wage) that is above the level determined in the market. Note that the first of these involves price control in product markets, while the second concerns price control in a resource market. While market outcomes are similar, each type of price control has different consequences for the economy and stakeholders. We will therefore consider each one separately.

A **price floor** is a **minimum price** set *below* the equilibrium price, in order to provide income support to farmers or to increase the wages of low-skilled workers.

To avoid getting confused about which is which, note that the position of a price floor and a price ceiling in relation to the equilibrium price is always *the opposite* of the floor and ceiling of a room. The *price floor is above* and the *price ceiling is below*.

Consequences of price floors

Consequences for markets

Farmers' incomes in many countries, resulting from the sale of their products in free markets, are often unstable or too low. Some important reasons for both instability and low incomes were considered in Chapter 3. Unstable incomes arise from unstable agricultural product prices, which are due to low price elasticities of demand and low price elasticities of supply for agricultural products. Low farmer incomes may arise from low income elasticities of demand.

One method governments use to support farmers' incomes is to set price floors for certain agricultural products, the objective being to raise the price above their equilibrium market price. The consequences are explained below.

Surpluses

Figure 4.6 illustrates the market for an agricultural product with a price floor, P_f, set above the equilibrium price, P_e. The price floor results in a larger quantity supplied, Q_s, than the quantity supplied at market equilibrium, Q_e. In addition, the price floor, P_f, leads to a smaller quantity demanded and purchased than at the equilibrium price: the quantity consumers want to buy at P_f is Q_d, which is smaller than the quantity Q_e that they bought at price P_e.

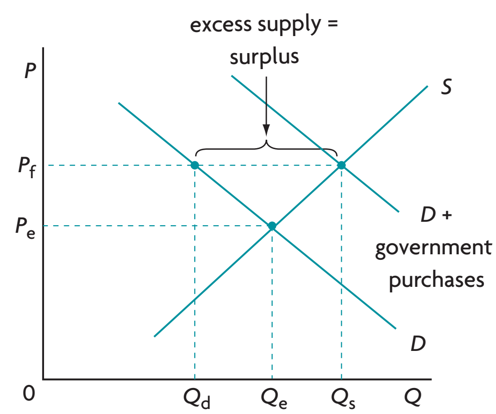


Figure 4.6: An agricultural product market with price floor and government purchases of the surplus

A price floor results in disequilibrium where there is a surplus (excess supply). A common practice is for the government to buy the excess supply, and this causes the demand curve for the product to shift to the right to the new demand curve '*D* plus government purchases'. By buying up the excess supply, the government is able to maintain the price floor at P_f.

It should be noted that if the government did not buy the surplus, the price would fall back to its equilibrium level. The reason is that farmers would have excess output with no buyers, and so would have to lower the price in order to be able to sell all the surplus.

Government measures to dispose of surpluses

The government must make a decision about what to do with the surplus (excess supply) it purchases. One option is to store it, giving rise to additional costs for storage above the costs of the purchase. Another method is to export the surplus (sell it abroad); this often requires granting a subsidy (this is money given to producers, to be discussed later in this chapter) to lower the price of the good since foreign countries would not want to buy it at the high price. Clearly, subsidies involve additional costs for the government. In general, any course chosen by the government to get rid of the surpluses is problematic.

Firm inefficiency

Higher than equilibrium product prices can lead to inefficient production; inefficient firms with high costs of production do not face incentives to cut costs by using more efficient production methods because the high price offers them protection against lower-cost competitors. This leads to inefficiency.

Overallocation of resources to the production of the good and allocative inefficiency

Too many resources are allocated to the production of the good, resulting in a larger than optimum (or 'best') quantity produced. Whereas the optimum quantity is Q_e, actually Q_s is produced.

Negative welfare impacts

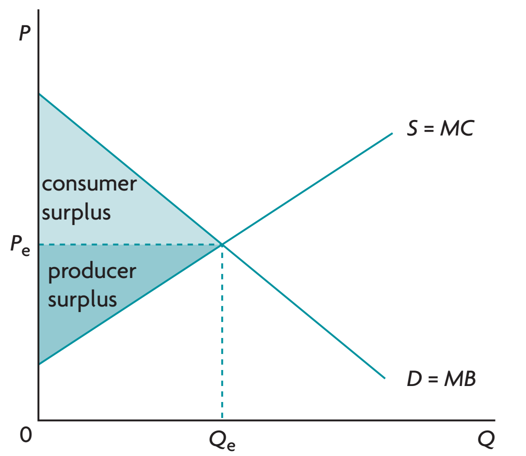
In Figure 4.7(a), price P_e and quantity Q_e represent market equilibrium with no price floor, and where social surplus is maximum. In part (b), when there is no price floor, consumer surplus is given by a + b + c and producer surplus by d + e. Also, $MB = MC$.

After a price floor, P_f, is imposed, consumer surplus becomes the area under the demand curve and above P_f, up to the quantity consumers buy, Q_d, and so falls to a. Producer surplus becomes the area above the supply curve and below P_f, up to the quantity produced, Q_s, and so becomes d + e + b + c + f. This means that the sum of consumer plus producer surplus *increases* by the area f after the price floor is imposed. This happens because producers gain the area b and c lost by consumers, and in addition gain f.

On the other hand, government spending to buy the excess supply is equal to the price paid per unit, P_f, times the surplus quantity it purchases: $P_f \times (Q_s - Q_d)$, corresponding to the rectangle outlined in bold. Since government spending is financed out of taxes with alternative uses (opportunity costs), government spending to maintain the price floor involves losses for society.

Therefore, there is a gain in surplus of f and a loss equal to the rectangle shown in bold. Subtracting the loss from the gain we are left with the blue shaded area, which is welfare loss, representing loss of benefits due to allocative inefficiency caused by overallocation of resources to the production of the good. This is also shown by $MB < MC$ at the point of production, Q_s, indicating that society would be better off if less of the good were produced.

a Consumer and producer surplus in a competitive free market: maximum social surplus



b Welfare impacts of a price floor (minimum price)

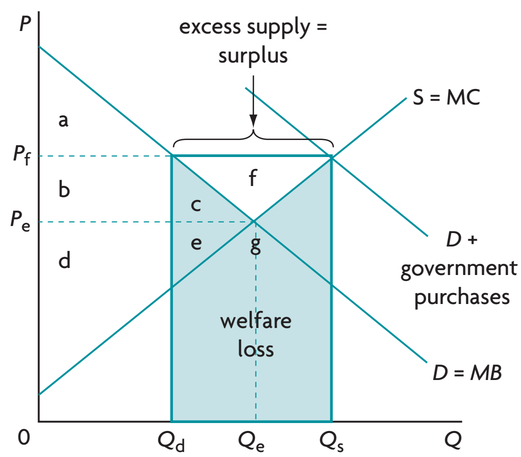


Figure 4.7: Welfare impacts of a price floor (minimum price) for agricultural products and government purchases of the surplus

Another way to see that the blue shaded area is the welfare loss is to note simply that the price floor results in a gain of f and a loss of c + e + f + g, or the bold rectangle of government spending. And so the net loss is simply that part of government spending that is not gained. In other words the net change is f − (c + e + f + g) = − (c + e + g).

A price floor creates welfare loss, indicating that the price floor introduces allocative inefficiency due to an overallocation of resources to the production of the good, shown by $Q_s > Q_e$. Also $MB < MC$, indicating that society is getting too much of the good.

Consequences for stakeholders

Consumers

Consumers are worse off, as they must now pay a higher price for the good ($P_f > P_e$), while they buy a smaller quantity of it ($Q_d < Q_e$). This is clear also from their loss of some consumer surplus.

Producers

Producers gain as they receive a higher price and produce a larger quantity, and since the government buys up the surplus, they increase their revenues from $P_e \times Q_e$ to $P_f \times Q_s$. Remember, this is the main rationale of agricultural price floors. Also, producers become protected against low-cost competition and do not face as strong incentives to become efficient producers; they are therefore less likely to go out of business if they are producing inefficiently (with higher costs).

Workers

Workers are likely to gain as employment increases on account of greater production of the good.

Government

When the government buys the excess supply, this is a burden on its budget, resulting in less government funds to spend on other desirable activities in the economy. The costs to the government are paid for out of taxes (and therefore by taxpayers). In addition, there are further costs of storing the surplus or subsidising it for export (sale to other countries).

Stakeholders in other countries

The European Union, the United States and many other more developed countries rely on price floors for agricultural products to support their farmers.

The surpluses are sometimes exported (sold to other countries), leading to lower world prices due to the extra supply made available in world markets. Countries that do not have price supports are forced to sell their agricultural products at low world prices. The low prices in these countries signal to local farmers that they should cut back on their production, resulting in an underallocation of resources to these products. These events often work against the interests of less developed countries (this topic will be discussed in Chapter 19).

Overall, a global misallocation of resources can result in a waste of resources, as price floors cause high-cost producers to produce more and low-cost producers to produce less than the social optimum.

TEST YOUR UNDERSTANDING 4.3

1. Define a price floor, and providing examples, explain some reasons why governments impose them.
2. Draw a diagram illustrating a price floor that is imposed in a product market, and analyse its effects on market outcomes (price, quantity demanded, quantity supplied, market disequilibrium, excess supply, firm inefficiency, allocative inefficiency, welfare loss).
3. Identify some measures governments can take to dispose of surpluses that result from the imposition of a price floor in an agricultural product market. Comment on the problems associated with these measures.
4. Assuming a price floor is imposed in a market for an agricultural product, and that the government purchases the entire excess supply that results in order to maintain the price:
   a. draw a diagram illustrating welfare loss, and
   b. comment on the relationship between marginal benefits and marginal costs in the new equilibrium and what it reveals about allocative efficiency (or inefficiency).
5. Examine the consequences for different stakeholders of a price floor for an agricultural product whose excess supply is purchased by the government.

Calculating effects of price floors on stakeholders and social welfare (HL only)

Figure 4.8 provides a numerical example of a price floor on an agricultural product. At equilibrium, price is equal to £20 and quantity is equal to 60 000 kg per week. When a price floor is imposed at P_f = £25, quantity demanded is Q_d = 40 000 kg per week and quantity supplied is Q_s = 80 000 kg per week.

Surplus (excess supply)

The surplus, or excess supply is equal to $Q_s - Q_d$, which in this case is 80 000 – 40 000 = 40 000 kg per week.

Change in consumer expenditure

Consumer expenditure is given by the price per kg of the good times the number of kg purchased per week. At equilibrium, before the price floor, consumers spend $P_e \times Q_e$ = £20 × 60 000 kg = £1.2 million per week. After the price floor is imposed, consumers spend $P_f \times Q_d$ = £25 × 40 000 kg = £1 million per week. Therefore, consumers spend £200 000 less on the good per week.

Change in producer revenue

Before the price floor, producer revenue is the same as consumer expenditure, since revenue is equal to price per kg times quantity sold, and both the price (P_e) and the quantity (Q_e) are the same for consumers and producers. Therefore producer revenue before the price floor is £1.2 million per week. Once the price floor is imposed and the government purchases the surplus (excess supply), firms receive revenues of $P_f \times Q_s$, and so producer revenue increases to £25 × 80 000 kg = £2 million per week. Therefore, the change is £800 000, or additional producer revenue of this amount per week.

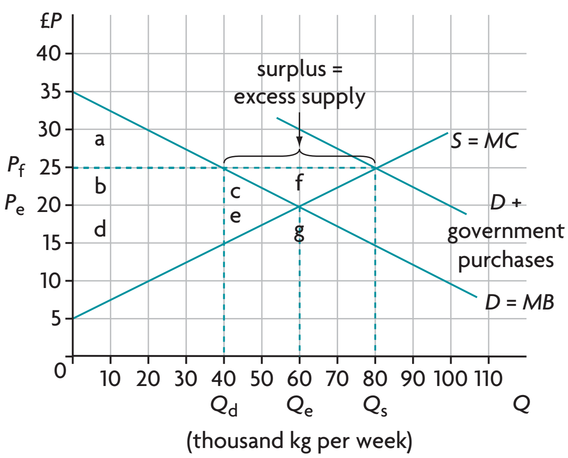


Figure 4.8: Calculating effects of a price floor on an agricultural product with government purchases of the surplus

Government expenditure

In order to purchase the excess supply of the agricultural product, the government spends an amount equal to the price of the good at the price floor times the number of kg purchased, or $P_f \times (Q_s - Q_d)$ = £25 × 40 000 = £1 million.

Note that government expenditure (£1 million) is equal to total producer revenue (£2 million) minus total consumer expenditure (£1 million) per week.

Change in consumer and producer surplus and welfare loss

Notice that all areas of consumer and producer surplus, both before and after the price floor, are the areas of triangles.

In this case, as we know, the formula for calculating consumer surplus is:

Consumer surplus =

$$\frac{(P \text{ intercept of D curve minus } P \text{ of consumers}) \times Q \text{ purchased}}{2}$$

Therefore, using the information in Figure 4.8, consumer surplus before the price floor is:

$$\text{Initial consumer surplus } \frac{(35-20)\times 60\,000}{2} = £450\,000$$

After the price floor is imposed consumer surplus is:

$$\text{Final consumer surplus} = \frac{(35-25)\times 40\,000}{2} = \frac{10\times 40\,000}{2}$$

$$= £200\,000$$

Therefore, consumer surplus *decreased* by £450 000 – £200 000 = £250 000.

Producer surplus is given by:

Producer surplus =

$$\frac{(P \text{ of producers minus } P \text{ intercept of } S \text{ curve}) \times Q \text{ sold}}{2}$$

Before the price floor producer surplus is:

$$\text{Initial producer surplus} = \frac{(20-5)\times 60\,000}{2} = \frac{15\times 60\,000}{2}$$

$$= £450\,000$$

After the price floor is imposed producer surplus is:

$$\text{Final producer surplus} = \frac{(25-5)\times 80\,000}{2} = \frac{20\times 80\,000}{2}$$

$$= £800\,000$$

Therefore, producer surplus *increased* by £800 000 – £450 000 = £350 000.

The welfare loss is given by the shaded area in Figure 4.7, which according to Figure 4.8 is the rectangle of government spending minus area f.

We have found above that government spending on the subsidy is £1 million.

$$\text{Area f} = \frac{(25-20)\times(80\,000-40\,000)}{2} = \frac{(5\times 40\,000)}{2}$$

$$= £100\,000$$

Therefore, welfare loss = £1 million – £100 000 = £900 000.

TEST YOUR UNDERSTANDING 4.4

1 In the example of the market illustrated in Figure 4.8, what would be the effect of a price floor set at £15?

2 The diagram below shows a price floor of \$50 that has been set for good Y. Calculate:

a the surplus (excess supply),

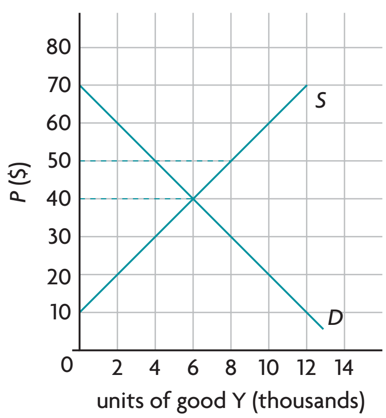
b the change in consumer expenditure,

c the change in producer revenue,

d the change in consumer surplus,

e the change in producer surplus, and

f welfare loss.



Minimum wages

Impacts of minimum wages on market outcomes

Many countries around the world have **minimum wage** laws that determine the minimum price of labour (the wage rate) that an employer (a firm) must pay. The objective is to guarantee an adequate income to low-income workers, who tend to be mostly unskilled. (The market-determined wages of skilled workers are usually higher than the minimum wage.) Figure 4.9 shows the market for labour. The demand for labour curve shows the quantity of labour that firms are willing and able to hire at each wage, and the supply of labour curve shows the quantity of labour that workers supply at each wage. Supply and demand determine the equilibrium 'price' of labour, which is the wage, W_e, where the quantity of labour demanded is equal to the quantity of labour supplied, Q_e.

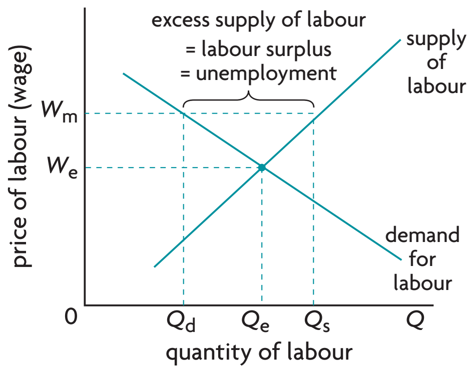


Figure 4.9: Labour market with minimum wage (price floor)

The minimum wage, W_m, lies above the equilibrium wage, W_e. Therefore, at W_m, the quantity of labour supplied, Q_s, is larger than the quantity of labour supplied when the labour market is in equilibrium (Q_e). The quantity of labour demanded, Q_d, is less than the quantity demanded at equilibrium, Q_e. There results a surplus of labour in the market equal to the difference between Q_s and Q_d. The labour market does not clear when there is a minimum wage.

Consequences of minimum wages for the economy

- ***Labour surplus (excess supply) and unemployment***

The imposition of a minimum wage in the labour market creates a surplus of labour equal to $Q_s - Q_d$ in Figure 4.9, which is unemployment, as it corresponds to people who would like to work but are not employed. The unemployment is due partly to the decrease in quantity of labour demanded by firms (the difference between Q_e and Q_d) and partly to an increase in the quantity of labour supplied (the difference between Q_s and Q_e). This occurs because the higher wage makes work more attractive, causing a movement up the labour supply curve. This unemployment is likely to involve unskilled workers.

- ***Illegal workers at wages below the minimum wage***

Illegal employment of some workers at wages below the legal minimum may result; this often involves illegal immigrants who may be willing to supply their labour at very low wages.

- ***Misallocation of labour resources***

The minimum wage affects the allocation of *labour resources*, as it prevents the market from establishing a market-clearing price of labour. In Chapter 2, we saw how the wage acts as a signal and incentive to workers (the suppliers of labour) and firms (the demanders of labour) to determine the optimal allocation of labour resources. The imposition of a minimum wage changes these signals and incentives for unskilled labour, whose wage is affected by the price floor. Therefore, industries that rely heavily on unskilled workers are more likely to be affected, and will hire less unskilled labour.

- ***Misallocation in product markets***

Firms relying heavily on unskilled workers experience an increase in their costs of production, leading to a leftward shift in their *product* supply curve (see Chapter 2), resulting in smaller quantities of output produced. Therefore, the misallocation of labour resources leads also to misallocation in product markets.

Consequences of minimum wages for various stakeholders

- ***Firms (employers of labour)***

Firms are worse off as they face higher costs of production due to the higher labour costs.

- ***Workers (suppliers of labour)***

The impacts on workers are mixed. Some gain because they receive a higher wage than previously ($W_m > W_e$), but some lose because they lose their job. Note that the workers who lose their job are those represented by $Q_e - Q_d$. This is not the full amount of unemployment created by the minimum wage, because the minimum wage leads to *additional* unemployment of $Q_s - Q_e$, since more workers supply their labour in the market when the wage increases.

- ***Consumers***

Consumers are negatively affected, because the increase in labour costs leads to a decrease in supply of products (a leftward shift in firm supply curves) causing higher product prices and lower quantities.

Price floors and minimum wages in the real world

Economists agree that price floors for agricultural products lead to surpluses (excess supplies) and are highly inefficient for the reasons discussed above. Yet they continue to be used in many countries because of strong political pressures exerted by farmers who claim to need these for income support.

The effects of minimum wages, on the other hand, are controversial, as it is questionable whether they lead to unemployment to the extent that economic theory predicts. There is agreement that if a minimum wage is set at a high level relative to the free market equilibrium wage, it is likely to create some unemployment. Yet a large and growing number of studies show that a minimum wage may have no effect or even a positive effect on *total* employment. Some firms respond to the minimum wage by maintaining the same number of workers but cutting non-wage benefits (such as paid holidays or sick leave). Also, it is possible that labour productivity (defined as the amount of output produced per worker) may increase due to the minimum wage, as workers feel motivated to work harder, with the result that some firms hire more unskilled labour in response to minimum wages.

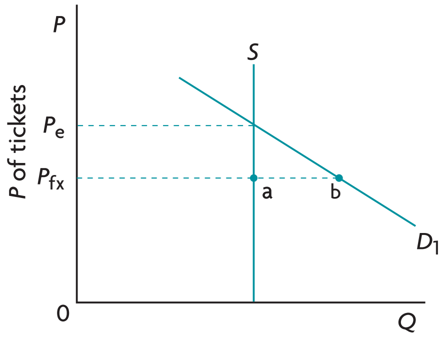
While the effects of minimum wages remain controversial, there is generally strong political support for their continued use on the grounds of greater equity in income distribution. In fact, the debate is not about whether or not there should be minimum wages, but rather what their level should be.

Setting fixed prices

Sometimes prices may be fixed at a particular level, such as with ticket prices for theatres, movies and sports events, where prices are usually fixed ahead of time by the organising body (which may be private or public), and so cannot increase or decrease according to supply and demand.

Figure 4.10 shows the market for tickets for a sports event. The supply curve is vertical because there is a fixed supply of tickets (due to a fixed number of seats; see Chapter 2). The ticket price is fixed at P_{fx} by the organising body. Figure 4.10(a) illustrates an event for which there is large demand, given by D_1. If the price could respond to market forces, it would rise to P_e, but since it is fixed at P_{fx} a shortage of tickets arises equal to the horizontal difference between points a and b. Figure 4.10(b) illustrates an event for which there is low demand, given by D_2. Here, the equilibrium price would have been P_e, however price is fixed at the higher level P_{fx}, resulting in a surplus of tickets equal to the horizontal difference between points c and d.

a Price fixing resulting in a shortage



b Price fixing resulting in a surplus

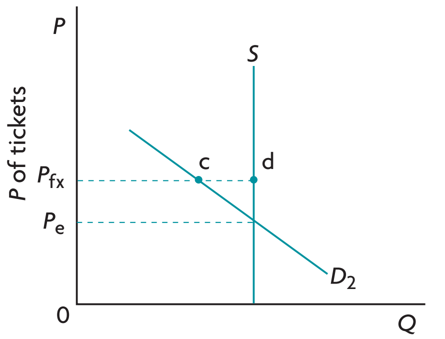


Figure 4.10: Price fixing and surpluses and shortages

TEST YOUR UNDERSTANDING 4.5

1. Define a minimum wage. Suggest reasons why many governments around the world impose them.
2. Draw a diagram illustrating the imposition of a minimum wage, and analyse its effects on market outcomes (the wage, quantity of labour demanded and supplied, market disequilibrium) and consequences for the labour market (unemployment of labour, illegal work, resource misallocation, welfare loss).
3. Suggest why, in practice, minimum wages may not lead to increased unemployment if they are not set too high.
4. Using diagrams show how excess demand or excess supply of tickets results when ticket prices are set at a level that is
   - a lower than equilibrium, and
   - b higher than equilibrium.

4.3 Indirect taxes

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the consequences of indirect taxes on markets and stakeholders (AO2)
- draw diagrams to illustrate the effects of indirect taxes on markets and stakeholders (AO4)
- evaluate the effects of indirect taxes on markets and stakeholders (AO3)
- calculate the effects of indirect taxes on markets and stakeholders (HL only) (AO4)

Introduction to indirect taxes

The meaning of indirect taxes

Indirect taxes are imposed on spending to buy goods and services. They are paid partly by consumers, but are paid to the government by producers (firms), and for this reason are called 'indirect'. There are two types of indirect taxes:

- *excise taxes*, imposed on particular goods and services, such as petrol (gasoline), cigarettes and alcohol
- *taxes on spending on all (or most) goods and services*, such as *general sales taxes* (used in the United States) and *value added taxes* (used in the European Union, Canada and many other countries).

Indirect taxes differ from **direct taxes**, involving payment of the tax by the taxpayers directly to the government (see Chapter 12).

In this chapter, we will study excise taxes.

Indirect taxes and the allocation of resources

Taxes have the effect of changing the allocation of resources. In Chapter 2 we learned that prices act as signals and incentives, which determine the pattern of resource allocation. Since indirect (excise) taxes are imposed on particular goods, they increase the price paid by consumers, causing consumers to reduce their spending on taxed goods. Excise taxes also lower the price received by producers, causing them to produce less. Therefore, by changing price signals and incentives, excise taxes affect the allocation of resources.

The interesting question is whether indirect taxes work to reduce or to increase allocative efficiency. The answer depends on the degree of allocative efficiency in the economy before the tax is imposed. If an economy begins with an efficient allocation of resources, the tax creates allocative inefficiency and a welfare loss. We will see how this happens below. In an economy with an inefficient resource allocation, indirect taxes potentially have the effect of improving resource allocation if they are designed to remove the source of allocative inefficiency. This will be studied in Chapter 5.

Why governments impose indirect taxes

Governments impose excise taxes for several reasons:

- **Indirect taxes are a source of government revenue.** Governments collect revenues from indirect taxes. As we know from Chapter 3, the lower the price elasticity of demand, the greater the government revenue generated.
- **Indirect taxes are a method to discourage consumption of goods that are harmful for the individual.** The consumption of certain goods is considered harmful for the individual (for example, cigarette smoking, excess alcohol consumption or gambling). Taxing these goods is likely to reduce their consumption. However, the extent to which these taxes are successful in reducing consumption depends on the price elasticity of demand; if it is low, an indirect tax will likely result in only a relatively small decrease in quantity demanded (see Chapter 5).
- **Indirect taxes can be used to redistribute income.** Some excise taxes focus on **luxury goods** (expensive cars, boats, furs, jewellery, and so on). The objective is to tax goods that can only be afforded by high-income earners. Payment of a tax on the purchase of these goods reduces after-tax income, thus narrowing differences with the incomes of lower-income earners.
- **Indirect taxes are a method to improve the allocation of resources (reduce allocative inefficiencies) by correcting negative externalities.** If there are market imperfections (in the form of negative externalities), preventing the achievement of allocative efficiency, indirect taxes can be used to try to improve the allocation of resources. This topic will be discussed in Chapter 5.

In this chapter, we assume that the economy begins with allocative efficiency in order to see how the introduction of indirect taxes leads to allocative inefficiency.

Indirect taxes: impacts on market outcomes and consequences for stakeholders

Distinguishing between specific and *ad valorem* taxes

Indirect, excise taxes can be:

- *specific taxes*, a fixed amount of tax per unit of the good or service sold; for example, €5 per packet of cigarettes
- *ad valorem taxes*, a fixed percentage of the price of the good or service; in this case, the amount of tax increases as the price of the good or service increases.

In our study of indirect taxes, we will consider only specific taxes, in other words a specific amount of tax for each unit of the good sold.

When a tax is imposed on a good or service, it is paid to the government by the firm. This means that for every level of output the firm is willing and able to supply to the market, it must receive a price that is higher than the original price by the amount of the tax. This involves a shift of the supply curve upward by the amount of the tax, and is shown in Figure 4.11(a). The tax causes a parallel upward shift, because the tax is a fixed amount for each unit of output. Therefore, in Figure 4.11(a) S_2 is parallel to S_1 (Note that this is equivalent to a leftward shift of the supply curve, meaning that for each price, the firm is willing to supply less output; this equivalence is explained in 'Quantitative techniques' chapter in the digital coursebook).

Illustrating and analysing impacts of an indirect tax on market outcomes

The impacts of specific taxes on market outcomes are shown in Figure 4.11(b). The supply curves are the same as in Figure 4.11(a); a demand curve has also been added. The pre-tax equilibrium is determined by the intersection of the demand curve D and the supply curve S_1, so the price paid by consumers and received by producers is P^* and quantity demanded and supplied is Q^*. If the government imposes a specific tax on the good, the supply curve shifts upwards to $S_2 = S_1$ + tax. The demand curve remains constant at D since demand is not affected. The new market equilibrium is determined by the demand curve D and the new supply curve S_2, so the price paid by consumers increases to P_c, and the quantity purchased falls to Q_t. The amount of tax per unit of output is shown on the vertical axis by $P_c - P_p$, or the vertical difference between the two supply curves. Whereas producers receive from consumers P_c per unit, they must pay the government $P_c - P_p$ per unit (tax per unit). Therefore, P_p is the final price received by producers after payment of the tax.

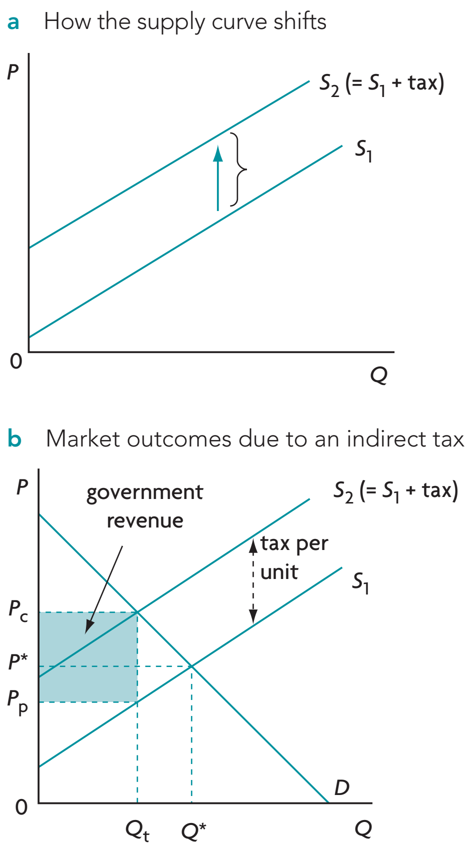


Figure 4.11: Supply curve shifts due to an indirect tax

The tax is said to 'drive a wedge' between the price P_c paid by consumers and the price P_p received by producers.

The market outcomes due to the tax are the following:

- equilibrium quantity produced and consumed falls from Q^* to Q_t
- equilibrium price increases from P^* to P_c, which is the price paid by consumers
- consumer expenditure on the good is given by the price of the good per unit times the quantity of units bought; it therefore changes from $P^* \times Q^*$ to $P_c \times Q_t$
- price received by the firm falls from P^* to P_p, which is $P_p = P_c$ – tax per unit
- the firm's revenue falls from $P^* \times Q^*$ to $P_p \times Q_t$
- the government receives tax revenue, given by $(P_c - P_p) \times Q_t$, or the amount of tax per unit times the number of units sold; this is the shaded area in Figure 4.12
- there is an underallocation of resources to the production of the good: Q_t is less than the free market quantity, Q^*.

Consequences of indirect taxes for various stakeholders

Consumers

Consumers are affected in two ways: by the increase in the price of the good (from P^* to P_c, shown in Figure 4.11(b)) and by the decrease in the quantity they buy (from Q^* to Q_t). Both these changes make them worse off, as they are now receiving less of the good and paying more for it.

Producers (firms)

Producers are affected in two ways: by the fall in the price they receive (from P^* to P_p), and by the fall in the quantity of output they sell (from Q^* to Q_t). These effects translate into a fall in their revenues, from $P^* \times Q^*$ before the tax to $P_p \times Q_t$. Firms are therefore worse off as a result of the tax.

The government

The government is the only stakeholder that gains, as it now has revenue equal to $(P_c - P_p) \times Q_t$ in Figure 4.11(b). This is positive for the government budget.

Workers

A lower amount of output, from Q^* to Q_t, means that fewer workers are needed to produce it; therefore, the tax may lead to some unemployment. Workers are worse off if they become unemployed.

Society as a whole: consumer and producer surplus and welfare loss

Society is worse off as a result of the tax, because there is an underallocation of resources to the production of the good ($Q_t < Q^*$). What happens to social surplus after the imposition of the tax? We can see this in Figure 4.12. Part (a) shows the maximum consumer plus producer surplus in a competitive free market that we are familiar with. The effects of the indirect tax can be seen in part (b), where consumer surplus becomes the shaded area under the demand curve and above P_c up to Q_t. Producer surplus becomes the shaded area above the supply curve S_1 and below P_p up to Q_t. A portion of consumer surplus became government tax revenue, and another portion was lost as triangle a. A portion of producer surplus also became government tax revenue, and another portion was lost as triangle b. Total government revenue can be seen in Figure 4.12(b).

Note that when identifying producer surplus after the imposition of an indirect tax, we always refer to the *initial supply curve*, which is S_1 in Figure 4.12b.

The consumer and producer surplus that is transformed into government tax revenue comes back to society in the form of government spending from the tax revenues. Therefore, the after-tax social surplus in Figure 4.12(b) is equal to after-tax consumer and producer surplus plus government revenue. However, after-tax social surplus is less than pre-tax social surplus by the amount of triangles a + b. The areas a + b represent social surplus that is completely lost, and is welfare loss.

In this case, welfare loss appears because the tax causes a smaller than optimum quantity to be produced: $Q_t < Q^*$. The tax has caused underproduction of the good relative to what is socially desirable, and an underallocation of resources, or allocative inefficiency.

Note that at the new point of production, Q_t, $MB > MC$, meaning that the benefits consumers receive from the last unit of the good they buy are greater than the marginal cost of producing it. Consumers would be better off if more of the good were produced (for an explanation see Chapter 2).

The imposition of an indirect tax results in reduced consumer and producer surplus, part of which is transformed into government revenue, and part of which is a welfare loss. The welfare loss in this case is the result of underallocation of resources to the production of the good (underproduction). This is also indicated by *MB* > *MC*: too little of the good is produced and consumed relative to the social optimum.

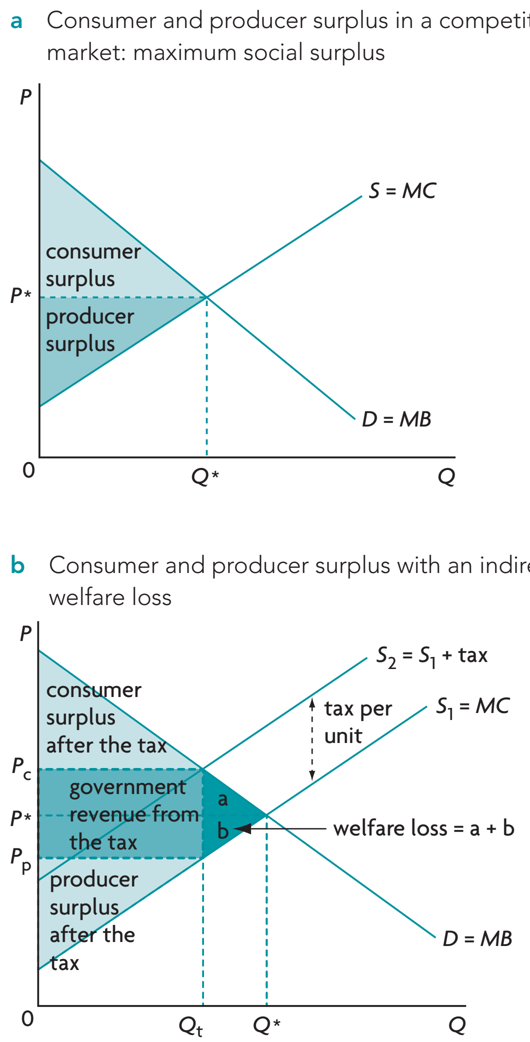


Figure 4.12: Effects of indirect taxes on consumer and producer surplus

TEST YOUR UNDERSTANDING 4.6

1 Describe what 'indirect' means in reference to 'indirect taxes'.

2 State some reasons why governments impose indirect taxes.

3 Outline how indirect taxes affect the allocation of resources.

4 Explain why in contrast to price controls, an indirect tax does not result in disequilibrium.

5 The government is considering imposing a €0.50 tax per litre of petrol (gasoline).

  a Draw a diagram for the gasoline market before the imposition of the tax, showing the price paid by consumers, the price received by producers and the quantity of petrol (gasoline) that is bought/sold.

  b Draw a diagram for the petrol (gasoline) market after the imposition of the tax, showing the price paid by consumers, the price received by producers and the quantity of petrol (gasoline) bought/sold.

6 For question (5):

  a analyse the impacts on the market of the tax on petrol (gasoline), and

  b discuss the consequences for stakeholders.

7 Using a diagram, show the effects of an indirect tax on consumer and producer surplus, as well as welfare loss. Explain why welfare loss arises as a result of a misallocation of resources.

Calculating the effects of indirect taxes on stakeholders and social welfare (HL only)

Suppose the government imposes an indirect (excise) tax on stromples of €6 per unit. This means that the supply curve will shift upward by €6 for each level of output Q.

How to graph the new supply curve, S_2

In Figure 4.13, the new supply curve, $S_2 = S_1$ + tax, lies €6 above the initial supply curve, S_1. We can count €6 upward along the vertical axis from the P intercept of S_1, and then draw a line parallel to S_1 from this new P intercept. This is gives us the new supply curve, S_2. For any Q, the vertical difference between the two supply curves is €6, which is the tax per unit of output.

How to find the new price paid by consumers, the price received by producers and the quantity bought and sold, following the imposition of the tax

After the tax is imposed, the demand curve D and the new supply curve, S_2, determine a new equilibrium price, which is P_c or the price paid by consumers, a new equilibrium quantity, Q_t, and P_p or the price received by producers ($P_p = P_c$ – tax per unit) (see also Figure 4.11(b)).

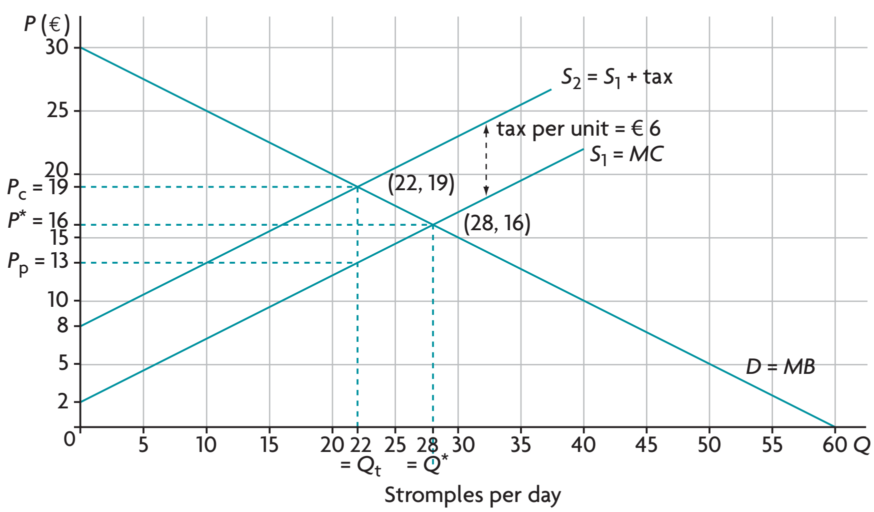


Figure 4.13: Demand and supply with indirect taxes

Therefore, at the new after-tax equilibrium, the price paid by consumers is P_c = € 19, the equilibrium quantity of stromples demanded and supplied is 22 units per day, i.e. $Q_t = 22$, and the price received by producers is $P_p = P_c$ – tax per unit = €19 – €6 = €13. These results are shown in Figure 4.13.

We have found that the price paid by consumers has increased from €16 to €19, the price received by producers has fallen from €16 to €13, and the quantity produced and consumed has fallen from 28 units to 22 units.

We now want to use this price and quantity information, together with the graph in Figure 4.13, to calculate the following: consumer expenditure, producer revenue, government revenue, consumer surplus and producer surplus.

Consumer expenditure

Consumer expenditure is given by the price paid per unit of stromples times the number of stromples purchased. Therefore, before the tax, consumers spent $P^* \times Q^*$ = €16 × 28 units = €448 per day; after the tax was imposed, consumers spent $P_c \times Q_t$ = €19 × 22 units = €418 per day. Therefore consumer expenditure fell by €30 per day (=€448 – €418).

Producer revenue

Producer revenue is given by the price received per unit of stromples times the number of stromples sold. Therefore, before the tax, producer revenue was $P^* \times Q^*$ = €16 × 28 units = €448 per day, which is the same as what consumers spent; firm revenue was exactly equal to consumer expenditure. After the tax was imposed, firm revenue fell to $P_p \times Q_t$ = €13 × 22 units = €286 per day.

Producer revenue fell by €162 per day (= €448 – €286). Firm revenue is now less than consumer expenditure.

Government revenue

Government revenue can be calculated in two ways:

a It is equal to tax per unit ($P_c - P_p$) times the number of units sold, Q_t, and is therefore €6 × 22 stromples = €132.

b It is also equal to the difference between consumer expenditure and producer revenue after the tax: €418 – €286 = €132.

Calculating the effects of indirect taxes on consumer surplus, producer surplus and welfare loss

As we have seen above, consumer and producer surplus both before and after the indirect tax are the areas of triangles, therefore we use the familiar triangle formula to calculate them.

Figure 4.14 is the same as Figure 4.12, but also shows P and Q values, which are the same values as in Figure 4.13. In part (a), consumer surplus is the shaded area under the demand curve and above $P^* = 16$, up to $Q^* = 28$. In part (b), it is the shaded area under the demand curve and above $P_c = 19$, up to $Q_t = 22$.

As you may recall, consumer surplus is:

Consumer surplus =

$$\frac{(P \text{ intercept of D curve minus } P \text{ of consumers}) \times Q \text{ purchased}}{2}$$

Therefore, consumer surplus before the tax is:

$$\frac{(30 - P^*) \times Q^*}{2} = \frac{(30-16) \times 28}{2} = \frac{14 \times 28}{2} = \frac{392}{2} = €196$$

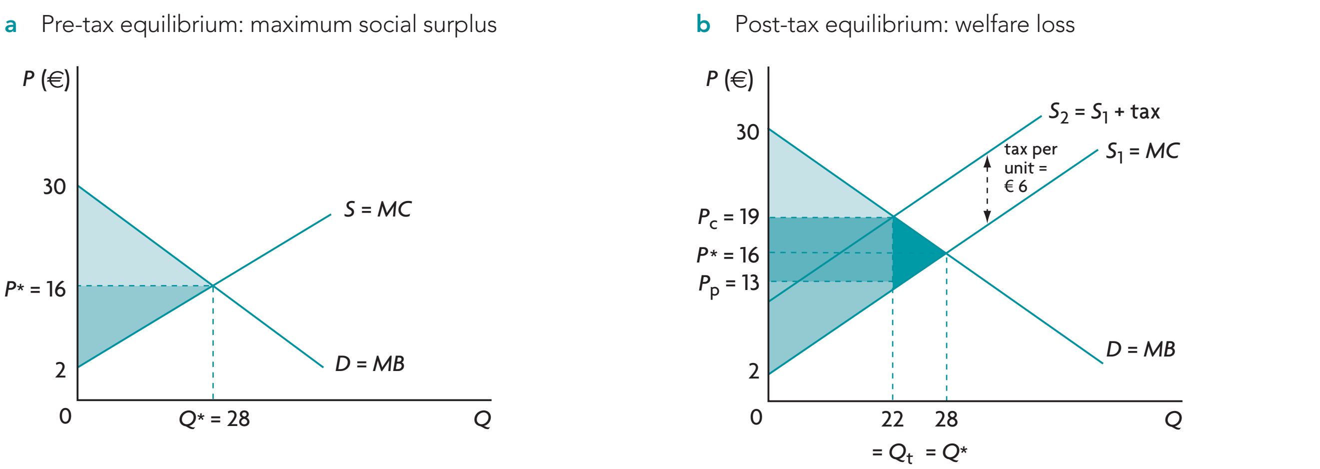


Figure 4.14: Calculating consumer and producer surplus before and after an indirect tax

Consumer surplus after the tax is:

$$\frac{(30-P_c)\times Q_t}{2}=\frac{(30-19)\times 22}{2}=\frac{11\times 22}{2}=\frac{242}{2}=€121$$

In the case of producer surplus, you may recall that it is:

Producer surplus =

$$\frac{(P \text{ of producers minus } P \text{ intercept of } S_1 \text{ curve})\times Q \text{ sold}}{2}$$

In Figure 4.14(a), producer surplus is the area above the supply curve S and below $P^* = 16$, up to $Q^* = 28$. In Figure 4.14(b), it is the area above the supply curve S_1 and below $P_p = 13$, up to $Q_t = 22$.

To calculate producer surplus, we can think of it as half the area of the rectangle whose one side equals the price received by producers minus the P intercept of the initial supply curve, S_1, and whose other side equals the number of units sold:

Producer surplus =

$$\frac{(P \text{ of producers minus } P \text{ intercept of } S_1 \text{ curve})\times Q \text{ sold}}{2}$$

Therefore, producer surplus before the tax is:

$$\frac{(P^*-2)*Q^*}{2}=\frac{(16-2)\times 28}{2}=\frac{14\times 28}{2}=\frac{392}{2}=€196$$

Producer surplus after the tax is:

$$\frac{(P_p-2)\times Q_t}{2}=\frac{(13-2)\times 22}{2}=\frac{11\times 22}{2}=\frac{242}{2}=€121$$

Note that to calculate producer surplus after the tax has been imposed, we use the *initial supply curve*, S_1.

(You may have noticed that consumer and producer surplus are equal to each other, both before and after the tax; *this is coincidental, as they need not be equal to each other*.)

The welfare loss can be found by taking the pre-tax sum of consumer and producer surplus (total social surplus), and subtracting from that the post-tax sum of benefits (post-tax consumer surplus, producer surplus and tax revenue): 196 + 196 – (121 + 121 + 132) = 18.

This is also equal to the area of the triangle:

$$\frac{(P_c-P_p)(Q^*-Q_t)}{2}=\frac{(19-13)(28-22)}{2}=\frac{6\times 6}{2}=€18$$

TEST YOUR UNDERSTANDING 4.7

1 Using the concept of allocative efficiency explain why an indirect tax creates welfare loss.

2 In the market for good zeta, the P intercept of the demand curve is at the point where Q = 0 and P = 7, and the P intercept of the supply curve is at the point where Q = 0 and P = 1. The point of intersection of the demand curve and the supply curve at free market equilibrium is at the point where Q = 6 and P = 4.

   a Draw the demand and supply curves, and identify the equilibrium price and quantity.

   b Suppose that price is measured in \$, and quantity of zeta in tonnes per day, and that a tax of \$2 per tonne is imposed; draw the new supply curve and identify the price paid by consumers, the price received by producers and the new equilibrium quantity.

   c Explain why the increase in price paid by consumers is smaller than the amount of tax per unit.

   d Using your results, calculate the change in consumer expenditure, the change in firm revenue, government revenue, the change in consumer surplus, the change in producer surplus and welfare loss.

   e Identify, in your diagram, the areas that correspond to government revenue, welfare loss and after-tax consumer and producer surplus.

   f Outline how the relationship between marginal benefit and marginal cost at the new (after-tax) equilibrium relates to allocative efficiency (or inefficiency).

Tax incidence and price elasticities of demand and supply (Supplementary material)

When a good is taxed, part of the tax is paid by consumers and part by producers; therefore the tax burden is shared between the two. If you are interested in seeing how the tax is shared you can read about it in the digital coursebook as Supplementary material.

4.4 Subsidies

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the consequences of subsidies on markets and stakeholders (AO2)
- draw diagrams to illustrate the effects of subsidies on markets and stakeholders (AO4)
- evaluate the effects of subsidies on markets and stakeholders (AO3)
- calculate the effects of subsidies on markets and stakeholders (HL only) (AO4)

Introduction to subsidies

The meaning of subsidies

A **subsidy**, in a general sense, refers to assistance by the government to individuals or groups of individuals, such as firms, consumers, industries or sectors of an economy. Subsidies may take the form of direct cash payments or other forms of assistance such as low-interest or interest-free loans (for example, to students, to low-income consumers for the purchase of goods and services such as housing, or to firms needing assistance), the provision of goods and services by the government at below-market prices; tax relief (i.e. paying lower or no taxes); and others.

In this section, we will consider only subsidies consisting of payments by the government to firms. Such payments are usually a fixed amount per unit of output, and are therefore *specific subsidies*.

Subsidies and the allocation of resources

Subsidies, like taxes, have the effect of changing the allocation of resources because they affect relative prices, thus changing the signals and incentives prices convey. A subsidy granted to a firm (or group of firms) has the effect of increasing the price received by producers, causing them to produce more, and lowering the price paid by consumers, causing them to buy more. Therefore, the allocation of resources changes and results in greater production and consumption than in the free market.

As with indirect taxes, we are interested in seeing whether the granting of a subsidy improves or worsens the allocation of resources. Here, too, the answer depends on the degree of allocative efficiency in the market before the subsidy. In an economy where resources are allocated efficiently, a subsidy introduces allocative inefficiency and welfare losses. This will be the topic of this section. But if the economy begins with allocative inefficiency (due to market imperfections), then a subsidy can work to improve the allocation of resources if it is designed to correct the source of the inefficiency. This will be examined in Chapter 6.

Why governments grant subsidies

There are several reasons why governments grant subsidies to firms:

- **Subsidies can be used to increase revenues (and hence incomes) of producers.** Subsidies have the effect of increasing the revenues of producers. Therefore, governments often grant subsidies to particular producers whose revenues (and therefore incomes) they would like to support.
- **Subsidies can be used to make certain goods (necessities) affordable to low-income consumers.** Subsidies have the effect of lowering the price of the good that is paid by consumers, thus making the good more affordable. For example, a government may wish to make a food staple (such as bread or rice) more affordable to low-income earners. It can do so by granting a subsidy to producers of the good.
- **Subsidies can be used to encourage production and consumption of particular goods and services that are believed to be desirable for consumers.** A subsidy has the effect of increasing the quantity of a good produced and consumed. If a government wishes to encourage consumption of a good because it is considered to be desirable (for example, education, vaccinations), it can use a subsidy to achieve this.
- **Subsidies can be used to support the growth of particular industries in an economy.** Since subsidies have the effect of increasing the quantity of output produced, if granted to firms in a particular industry, they support the growth of that industry. For example, subsidies to the solar industry are intended to promote the growth of solar power.

Other examples include chemicals, textiles, steel, fossil fuels and many more.

- **Subsidies can be used to encourage exports of particular goods.** Since subsidies lower the price paid by consumers, they are sometimes granted on goods that are exported (sold to other countries), since lower export prices increase the quantity of exports (see Chapter 14).
- **Subsidies are a method to improve the allocation of resources (reduce allocative inefficiencies) by correcting positive externalities.** It was noted above that market imperfections prevent the achievement of allocative efficiency; in some cases (such as when there are positive externalities), it may be possible to use subsidies to improve allocative efficiency (see Chapter 6).

Subsidies are a controversial topic in economics because they are very extensive and are often designed to achieve certain objectives that may not be consistent with other important objectives. For example, many countries grant subsidies to fossil fuels, which run contrary to objectives of sustainable development and which also contradict the objectives of other subsidies intended to support the growth of alternative energy. Fossil-fuel subsidies are known as 'perverse subsidies'. Subsidies for agriculture and exports are also highly controversial (see Chapter 14).

Subsidies: impacts on market outcomes and consequences for stakeholders

Illustrating and analysing impacts of subsidies on market outcomes

In Figure 4.15, the initial, pre-subsidy equilibrium is determined by the intersection of the demand curve *D* and the supply curve S_1, giving rise to equilibrium price P^* paid by consumers and received by producers, and equilibrium quantity Q^*. Now the government grants a subsidy consisting of a payment to the firm of a fixed amount for each unit of output sold. This means that for each unit of output the firm is willing and able to produce, it receives a lower price than the original by the amount of the subsidy; this produces a downward, parallel shift of the supply curve by the amount of the subsidy, to the new curve $S_2 = S_1 -$ subsidy. (The reason we subtract the subsidy is that it works to *decrease the firms' costs of production*, thus causing a downward shift of the S curve. It is the exact opposite of the indirect tax which is added to the S curve to cause an upward shift. (See 'Quantitative techniques' chapter in the digital coursebook). The demand curve remains constant at *D* since demand is not affected. The demand curve and new supply curve S_2 determine a new equilibrium, where price is P_c (the price paid by consumers) and the quantity produced and sold increases to Q_{sb}. Since the vertical difference between the two supply curves represents the subsidy per unit of output, the firm receives price P_p, which is equal to the price paid by the consumer, P_c, plus the subsidy per unit of output.

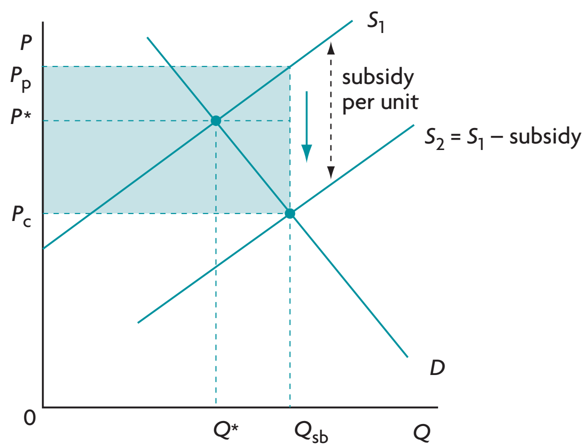


Figure 4.15: Impacts of subsidies on market outcomes

The market outcomes due to the subsidy are the following:

- equilibrium quantity produced and consumed increases from Q^* to Q_{sb}
- the equilibrium price falls from P^* to P_c; this is the price paid by consumers
- the price received by producers increases from P^* to P_p
- the amount of the subsidy is given by $(P_p - P_c) \times Q_{sb}$, or the amount of subsidy per unit multiplied by the number of units sold; this is the entire shaded area, and represents government spending to provide the subsidy
- there is an overallocation of resources to the production of the good: Q_{sb} is greater than the free market quantity, Q^*.

Consequences of subsidies for various stakeholders

Consumers

Consumers are affected by the fall in price of the good from P^* to P_c (Figure 4.15) and the increase in quantity purchased (from Q^* to Q_{sb}). Both these changes make them better off.

Producers

Producers are also better off, because they receive a higher price ($P_p > P^*$) and produce a larger quantity ($Q_{sb} > Q^*$), seen in Figure 4.15. The price and quantity effects translate into an increase in revenues. Before the granting of the subsidy, firms had revenues of $P^* \times Q^*$. Following the subsidy, firm revenues increase to $P_p \times Q_{sb}$.

The government

The government pays the subsidy, which is a burden on its budget. To obtain the revenues for the subsidy, the government may have to reduce expenditures elsewhere in the economy, or it may have to raise taxes, or it may have to run a budget deficit (government expenditures greater than tax revenues). Whatever the case, the impact on the government's budget is negative.

Workers

As output expands from Q^* to Q_{sb}, firms are likely to hire more workers to produce the extra output, therefore workers who find new jobs are better off.

Society as a whole: consumer and producer surplus, and welfare loss

Society as a whole is worse off because there is an overallocation of resources to the production of the good; $Q_{sb} > Q^*$. In addition, society is worse off because the higher price received by producers protects relatively inefficient ones, allowing them to continue to produce.

Figure 4.16 shows consumer and producer surplus before and after the subsidy. In part (a), at the free market equilibrium before the subsidy, social (consumer plus producer) surplus is maximum and $MB = MC$, indicating the achievement of allocative efficiency.

After the granting of the subsidy, both consumer surplus and producer surplus increase. In Figure 4.16(b), consumer surplus is now the area under the demand curve and above price P_c, up to output Q_{sb}. The *gain* in consumer surplus is shown by the shaded area labelled 'gain in consumer surplus'. Producer surplus becomes the area above the supply curve S_1 and below the price P_p, up to output Q_{sb}. The *gain* in producer surplus is shown by the shaded area labelled 'gain in producer surplus'.

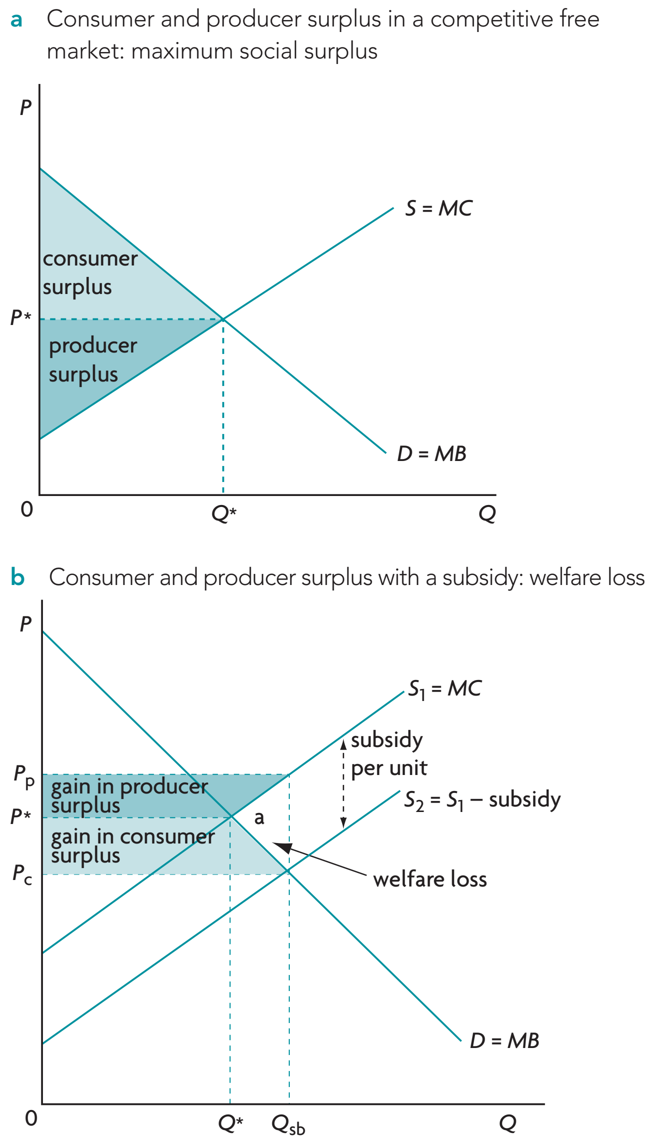


Figure 4.16: Effects of subsidies on consumer and producer surplus

Note that when identifying producer surplus after the granting of a subsidy, we always refer to the *initial supply curve*, which is S_1 in Figure 4.16.

At the same time that producers and consumers gain, the government loses because of the negative effects on its budget. The subsidy is paid for by taxes that have an opportunity cost (alternative uses that are sacrificed). As we know, government expenditure to provide the subsidy is $(P_p - P_c) \times Q_{sb}$. *This is* *exactly equal to the gains in consumer and producer surplus plus the triangle a*. Therefore, the social losses due to government spending are greater than the gains in consumer and producer surplus by the amount a. The area a is welfare loss representing lost benefits for society, caused by a larger than optimum quantity produced: $Q_{sb} > Q^*$. The subsidy has caused overproduction relative to what is socially desirable, and an overallocation of resources, or allocative inefficiency.

We can also see in Figure 4.16(b) that at Q_{sb}, $MB < MC$, meaning that the benefit consumers receive from the last unit of the good they buy is less than the marginal cost of producing it. Therefore, society would be better off if less of the good were produced.

The granting of a subsidy results in greater consumer and producer surplus; however, society loses due to government spending on the subsidy. Since the loss from government spending is greater than the gain in consumer and producer surplus, welfare loss results, reflecting allocative inefficiency, which in this case is due to overallocation of resources to the production of the good (overproduction). This is also illustrated by $MB < MC$: too much of the good is being produced and consumed relative to the social optimum.

Foreign producers

If the subsidy is granted on exports (goods sold to other countries), it lowers price and increases the quantity of exports. While this is positive for domestic producers, it is negative for the producers of other countries who may be unable to compete with the lower price of the subsidised goods. (This topic will be discussed in Chapters 14 and 18.)

TEST YOUR UNDERSTANDING 4.8

1 Define subsidies and providing examples, outline some reasons why governments grant them.

2 The government is considering granting a €0.50 subsidy per kilogram of cheese.

   a Draw a diagram for the cheese market before the granting of the subsidy, showing the price paid by consumers, the price received by producers and the quantity of cheese that is bought/sold.

   b Using the same diagram, show what happens in the cheese market after the granting of the subsidy, indicating the price paid by consumers, the price received by producers, the quantity of cheese bought/sold and government spending on the subsidy.

3 Considering question 2 above:

   a analyse the impacts on the market of the subsidy, and

   b discuss the consequences for stakeholders.

4 Using a diagram, show the increase in consumer and producer surplus that results from the granting of a subsidy. Explain why welfare loss arises even though both consumer and producer surplus increase. Show the welfare loss in your diagram.

REAL WORLD FOCUS 4.2

Farm subsidies in the United States

Farmers in the United States have been receiving subsidies for certain agricultural products (corn, wheat, soybeans, cotton, rice and others) since the Great Depression of the 1930s. In 1996, a law was passed to end subsidies and create a free market in agriculture. However, the free market in agriculture never materialised. Support to farmers costs tax payers over $20 billion a year. This support is justified by the common belief that the government is helping small farmers survive. Yet according to the US Department of Agriculture (the agriculture ministry), the largest and wealthiest farmers receive the bulk of

CONTINUED

farm subsidies. The reason is that subsidies are paid according to the amount of crop produced. Smaller farmers receive very small amounts, while farmers who cultivate fruits and vegetables receive no subsidies at all. In fact, many billionaires (who are the owners of very large agricultural companies) receive farm subsidies. According to the Environmental Working Group, 50 people in the Forbes list of 400 wealthiest Americans received subsidies in the period 1995 to 2014. Since 2008, the top ten recipients of subsidies *each* received an average of $18.2 million.

Source: *Adapted from 'Farm subsidies no illogical entitlements – next senator must push for overhaul of ag policy' in Lexington Herald-Leader, 11 July 2010. Chris Edwards, 'Reforming federal farm policies', Tax and budget bulletin no 82, Cato institute, 12 April 2018 'Mapping the US farm subsidy $1million club,' Adam Andrzejewski, Forbes, 14 August 2018*

Figure 4.17: Washington, D.C., USA. The National Farmers Union urges Congress to pass a farm bill

Applying your skills

1 Discuss the consequences of farm subsidies for agricultural markets and stakeholders.

2 Suggest reasons why the government continues to grant agricultural subsidies even though they are costly and unfair.

Calculating the effects of subsidies on market outcomes and social welfare (HL only)

If the government grants a subsidy on stomfles of $4 per kg, the supply curve shifts downward by $4 for each kg of output Q.

How to graph the new supply curve, S_2

In Figure 4.18, the new supply curve, $S_2 = S_1$ – subsidy, lies $4 below the initial supply curve, S_1. We can count $4 downward along the vertical axis from the P intercept of S_1, and then draw a line parallel to S_1 from this new P intercept. This gives us the new supply curve, S_2. For any Q, the vertical difference between the two supply curves is $4, which is the subsidy per unit of output.

Therefore, at the new equilibrium, the price paid by consumers is P_c = $18, the equilibrium quantity is Q_{sb} = 24 kg, and the price received by producers is $P_p = P_c$ + subsidy per unit = $18 + $4 = $22. These results are shown in Figure 4.18.

We have found that the price paid by consumers has fallen from $20 to $18 per kg, the price received by producers has increased from $20 to $22 per kg, and the quantity produced and consumed has increased from 20 kg to 24 kg.

We will now use this price and quantity information together with the graph in Figure 4.18 to calculate consumer expenditure, producer revenue, government expenditure, consumer surplus, producer surplus and welfare loss.

Consumer expenditure

Consumer expenditure equals the price paid per kg of stomfles times the number of kg purchased. Therefore, before the subsidy, consumers spent $P^* \times Q^*$ = $20 × 20 kg = $400 per day; after the subsidy, consumers spent $P_c \times Q_{sb}$ = $18 × 24 kg = $432 per day. Therefore consumer expenditure increased by $32 per day (= $432 – $400).

Producer revenue

Producer revenue is given by the price received per kg of stomfles times the number of kg sold. Therefore, before

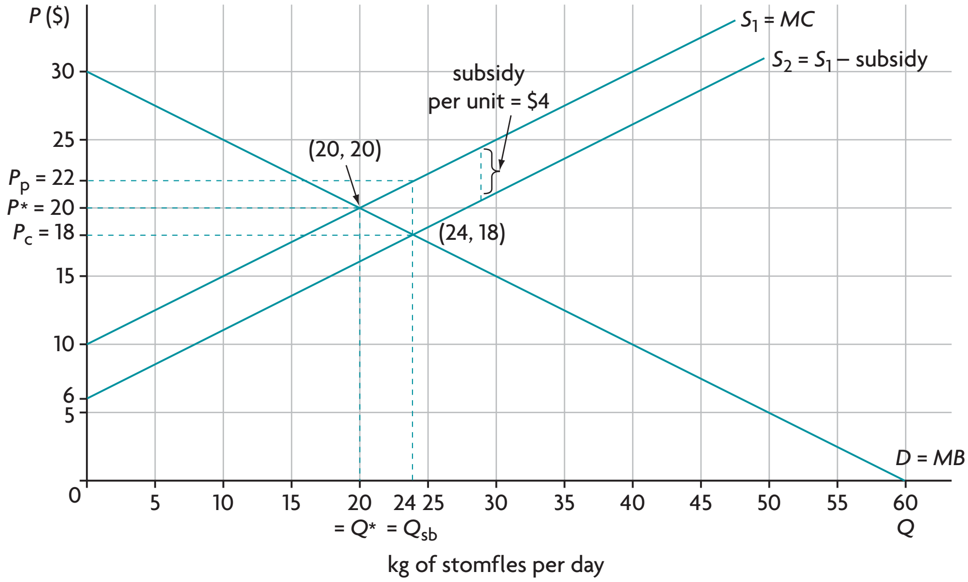


Figure 4.18: Demand and supply with subsidies

the subsidy was granted, producer revenue was $P^* \times Q^*$ = \$20 × 20 kg = \$400 per day, which is exactly the same as what consumers spent; firm revenue was exactly equal to consumer expenditure. After the subsidy was granted, producer revenue increased to $P_p \times Q_{sb}$ = \$22 × 24 kg = \$528 per day. Producer revenue increased by \$128 per day (= \$528 – \$400). Note that firm revenue is now more than consumer expenditure.

Government expenditure

Government expenditure on the subsidy can be calculated in two ways:

a It is equal to the subsidy per kg ($P_p - P_c$) times the number of kg sold (Q_{sb}), and is therefore \$4 × 24 kg = \$96 per day.

b It is also equal to the difference between producer revenue and consumer expenditure after the subsidy: \$528 – \$432 = \$96 per day.

Calculating the effects of subsidies on consumer and producer surplus and welfare loss

Consumer and producer surplus both before and after the subsidy are the areas of triangles, therefore we use the familiar triangle formula to calculate them.

Examining Figure 4.18, we can see that consumer surplus at the free market equilibrium (before the subsidy) is represented by the area under the demand curve and above price P^* = 20 up to quantity Q^* = 20. After the subsidy, it is the area under the demand curve and above price P_c = 18 up to quantity Q_{sb} = 24.

Consumer surplus =

$$\frac{(P \text{ intercept of D curve minus } P \text{ of consumers}) \times Q \text{ purchased}}{2}$$

Therefore, consumer surplus before the subsidy is:

$$\frac{(30 - P^*) \times Q^*}{2} = \frac{(30 - 20) \times 20}{2} = \frac{10 \times 20}{2} = \frac{200}{2} = \$100$$

Consumer surplus after the subsidy is:

$$\frac{(30 - P_c) \times Q_{sb}}{2} = \frac{(30 - 18) \times 24}{2} = \frac{12 \times 24}{2} = \frac{288}{2} = \$144$$

Therefore, consumer surplus increased by \$44 (= \$144 − \$100).

Producer surplus before the subsidy in Figure 4.18, is the area above the supply curve S_1 and below P^* = 20 up to Q^* = 20. After the subsidy, it is the area above the supply curve S_1 and below P_p = 22 up to Q_{sb} = 24.

Producer surplus is also calculated using the same method as with indirect taxes:

Producer surplus =

$$\frac{(P \text{ of producers minus } P \text{ intercept of } S_1 \text{ curve}) \times Q \text{ sold}}{2}$$

Therefore, producer surplus before the subsidy is:

$$\frac{(P^*-10)\times Q^*}{2}=\frac{(20-10)\times 20}{2}=\frac{10\times 20}{2}=\frac{200}{2}=\$100$$

Producer surplus after the subsidy is:

$$\frac{(P_p-10)\times Q_{sb}}{2}=\frac{(22-10)\times 24}{2}=\frac{12\times 24}{2}=\frac{288}{2}=\$144$$

Therefore producer surplus increased by \$44 (= \$144 − \$100).

> Note that to calculate producer surplus after the subsidy has been granted, we use the *initial supply curve*, S_1 (as we do also in the case of indirect taxes).

(The equality of producer and consumer and producer surplus, before and after the subsidy, is coincidental.)

Social surplus increased by the amount of the increase in consumer surplus plus the amount of the increase in producer surplus = \$44 + \$44 = \$88.

Welfare loss can be found by:

welfare loss =

$$\frac{(P_p-P_c)(Q_{sb}-Q^*)}{2}=\frac{(22-18)(24-20)}{2}=\frac{4\times 4}{2}=\$8$$

TEST YOUR UNDERSTANDING 4.9

1 Using the concept of allocative efficiency, explain why a subsidy creates welfare loss.

2 In the market for good alpha the *P* intercept of the demand curve is at the point where Q = 0 and P = 7 and the *P* intercept of the supply curve is at the point where Q = 0 and P = 1. The point of intersection of the demand curve and the supply curve at free market equilibrium is at the point where Q = 6 and *P* = 4.

   a Draw the demand and supply curves, and identify the equilibrium price and quantity.

   b Suppose that price is measured in £, and quantity in tonnes per day, and that a subsidy of £2 per tonne is granted. Draw the new supply curve, and find (through your graph) the price paid by consumers, the price received by producers and the new equilibrium quantity.

   c Using your results, calculate the change in consumer expenditure, the change in firm revenue, government expenditure, the change in consumer surplus, the change in producer surplus and welfare loss.

   d Identify, in your diagram, the areas that correspond to government expenditure, welfare loss, the increase in consumer surplus and the increase in producer surplus.

   e Outline how the relationship between marginal benefit and marginal cost at the new (after-subsidy) equilibrium relates to allocative efficiency (or inefficiency).

THEORY OF KNOWLEDGE 4.1

Allocative efficiency: is it really value-free?

Throughout this chapter, we have used the competitive market model, explained in Chapter 2, as the basis for making assessments about government intervention in the economy. According to this model, when there is competition in the sense of many buyers and sellers who act according to their best self-interest, and when market forces are free to determine equilibrium prices, a situation is reached where there is allocative efficiency and maximum social welfare. Scarce resources are allocated in the best possible way, producing the most of what people mostly want, and it is not possible to make anyone better off without making someone worse off, a condition called *Pareto optimality*.

CONTINUED

The concept of Pareto optimality emerged in the late 19th century after a period when economists were trying to make economics more scientific in its approach. 'Scientific' meant economics should get rid of any value judgements about things that 'ought to be' and base itself entirely on positive thinking (see Chapter 1). The famous classical economists of earlier times (Adam Smith, Thomas Robert Malthus, David Ricardo, John Stuart Mill, Karl Marx and many others) openly discussed their ideas about what ought to happen in society (normative ideas) together with their positive ideas of things that 'are' or 'will be'. Yet by the late 19th century, it was believed that a true science is value-free, and economists set out to imitate the methods of the natural sciences, especially physics.

The concept of Pareto optimality, developed by Vilfredo Pareto (an Italian engineer, sociologist, economist and philosopher), was welcomed as being truly free of normative aspects. It simply stated that under certain assumptions, numerous freely acting consumers and producers behaving according to their individual preferences, give rise to an outcome of maximum efficiency and maximum social welfare (defined as $MB = MC$ or maximum social surplus).

On the surface, this sounds like a value-free, positive statement. Yet in later years (and up to the present), many economists criticised it on the grounds that it is actually heavily based on normative ideas, and is therefore not value-free. The difficulty in detecting these values is that these are implicit; they are not explicit, or *openly stated* values.

Economists who question the value-free nature of Pareto optimality point out that the concept of 'welfare' is defined in relation to individual preferences of consumers and producers (decisions made on the basis of rational self-interest) that determine consumer and producer surplus. Individual preferences become the standard, or measuring stick, by which economists evaluate real-world situations and government intervention in markets. As we saw in this chapter, a free competitive market is 'best'; government intervention in this market reduces welfare.

But surely there are other definitions of welfare, such as equality of opportunity, freedom from hunger and disease, human rights, fairness and many more. All these are normative concepts, but then, by the same token, welfare defined as maximum social surplus is also a normative concept, on the basis of which a judgement is made about how well or how poorly the economy works.

Some economists take these ideas further, and argue that not only Pareto optimality but the body of microeconomics on which it rests is also not value-free. The idea that societies should pursue allocative efficiency follows from *the definition of economics as the social science that tries to determine how to make the best use of scarce resources*. Yet, remember from Chapter 1, all societies must answer three basic economic questions: *what/how much, how* and *for whom to produce*. The definition of economics covers only the first two of these and ignores the third. The reason for neglecting the *for whom to produce* question is that economists consider the first two questions to be part of positive thinking and the third a part of normative thinking (see Theory of knowledge 2.1 in Chapter 2). However, if efficiency (*what* and *how to produce*) leads to maximum welfare, where the definition of welfare is based on a value judgment, economists have come full circle and have based their so-called positive analysis on a value judgement.

In view of the above, some economists argue that the focus of government policies on efficiency, defined as Pareto optimality, diverts attention away from the problem of income distribution, and justifies government inaction in this area. What is the point of realising (or coming close to realising) Pareto optimality in the real world, if a large portion of the population is starving because they have no income?

CONTINUED

According to Gunnar Myrdal, a Swedish economist who won the Nobel Prize in 1974, the social sciences are inevitably based on values, but these values should be made explicit:

> The only way we can strive for 'objectivity' in theoretical analysis is to expose the valuations into full light, making them conscious, specific and explicit, and permit them to determine the theoretical research [...] there is nothing wrong, *per se*, with value-loaded concepts if they are clearly defined in terms of explicitly stated value premises.[1]

Thinking points

- In your view, is Pareto optimality value-free, or is it implicitly based on a value judgement?
- Does language in the expression 'maximum social welfare' convey values? Is it possible to have value-free language?
- Does the inevitable use of language in the pursuit of economic knowledge complicate the job of economists as social *scientists* in pursuit of value-free knowledge?
- Do you think the natural sciences are value-free?
- Do you agree with Myrdal's claim that it may be possible to reach 'objectivity' in theoretical social science by making values explicit?
- Do you think economics is, or ever can be, completely value-free?

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Find out if the city or area you live in uses rent controls. If it doesn't, search online and find a city or area where rent controls are in place. Research the effects of the rent controls and evaluate their usefulness. Consider subsidies on housing, which is alternative policy that could have been used to make housing affordable to people on low incomes. Compare and contrast the two policies and try to determine the advantages and disadvantages of each. Using your knowledge of economics, determine which of the two policies should be preferred. Justify your conclusion.

2 Find out if subsidies are granted in your country. If not, find a real-world example of one or more subsidies, research the stakeholders involved and the reasons why the subsidies are given. Evaluate the merits of the subsidies by considering the effects of the subsidies on markets and stakeholders.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

[1] G. Myrdal (1970) *Objectivity in Social Research*, Gerald Duckworth.

Chapter 5

Market failure and socially undesirable outcomes I:

Common pool resources and negative externalities

BEFORE YOU START

- Can you think of some goods or services that when consumed or produced cause harm to others even though they didn't consume or produce them?
- What do you think should be the role of government to reduce the harm caused by such harmful actions?

In this chapter we will discover why the market economy fails to achieve many of its promises. In particular we will see how some actions of individuals or groups of individuals may have negative impacts on others and on the environment. We will also study how government intervention can help markets overcome their shortcomings.

Photo: Borneo, Malaysia. Long tailed macaque monkey eating plastic bag in Bako National Park

5.1 The meaning of common pool resources

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the meaning of common pool resources in terms of (AO2)
  - the concepts of rivalry and non-excludability
  - the tragedy of the commons
- explain the meaning of unsustainable production (AO2)

Environmental problems can be studied by examining a special category of resources known as common pool resources.[1] **Common pool resources** are resources that are not owned by anyone, do not have a price and are available for anyone to use without payment or any other restriction. Examples include clean air, lakes, rivers, fish in the open seas, wildlife, hunting grounds, forests, biodiversity, the fertility of the soil that occurs in nature, open grazing land, the ozone layer, the stable global climate, and many more.

Understanding common pool resources

Common pool resources: rivalrous but non-excludable

Common pool resources differ from any other kind of resource or good, because they possess a special combination of characteristics: they are rivalrous and non-excludable. To understand what these terms mean, it is useful to consider the definition of private goods. A *private good* has two characteristics:

- It is **rivalrous**: its consumption by one person reduces its availability for someone else; for example, your computer, textbook, pencils and clothes are rivalrous, because when you buy them, another person cannot buy the same ones; most goods are rivalrous.
- It is **excludable**: it is possible to exclude people from using the good; exclusion is usually achieved by charging a price for the good; if someone is unwilling or unable to pay the price, he or she will not have the benefit of using it; most goods are excludable.

As noted above, most goods are rivalrous. Common pool resources are also rivalrous. If we use up clean air, there is less left over for use by others; when we catch fish in the open sea, there are fewer fish left over for others to catch; if we destroy the stability of the global climate, it will not be available for use by future generations.

Also, it was noted that most goods are excludable. However, open pool resources differ because *they have no price or any other means of excluding users; anyone can use them without payment or other restriction*; therefore they are **non-excludable**. Non-excludable means it is not possible to exclude someone from using a good or resource.

> Common pool resources are *rivalrous* but *non-excludable*. This combination poses serious threats to the environment. *Rivalry* means that the use of resources reduces their availability for others. *Non-excludability* means that the resources can be used abundantly without restrictions and therefore may be overused, degraded and depleted.

There is no end to examples of overuse, depletion and degradation of common pool resources. When factories, homes or cars use fossil fuels that emit pollutants into the atmosphere or into oceans, rivers and lakes, they 'overuse' a portion of these natural resources without paying for them. Some of these activities give rise to global warming, with likely devastating effects on agriculture, health and ecosystems; this involves 'overusing' the benefits provided by a stable global climate. When fish are overfished, the fishing industry uses up an excessive amount of the global stock of fish and disrupts the marine ecosystem. Similarly, when forests are cleared to create land for use in agriculture or for the sale of timber by the lumber industry, there

[1] It may be noted that in the previous syllabus these were known as *common access resources*. The term was changed as *common pool resources* is more commonly used.

are huge consequences in terms of loss of biodiversity and threats to wildlife, the ozone layer and the global climate. Land is being overgrazed because of excessive grazing; arable land is lost because of soil erosion and salinisation; wildlife is endangered because of the destruction of natural habitats due to the encroachment of settlers and agriculture. *In all these cases, common pool resources are used and overused, leading to serious environmental degradation and depletion.*

Common pool resources: The tragedy of the commons

The **tragedy of the commons** is a story about cattle that feed on a fertile pasture that is owned in common by a group of herders (cattle owners). In the beginning each herder had a small number of cattle and all the animals had plenty of space and grass on which to feed. As this was a profitable business, each herder began to increase the number of cattle grazing on the pasture. But as the cattle increased, after some time the pasture became overfilled with grazing cattle that had to increasingly compete with each other for food that was becoming more and more scarce. In the end the grass was all gone, the soil was eroded and the pasture could no longer be used for grazing.

This story allows us to better understand the concepts of rivalry and non-excludability. The fertile pasture is rivalrous because whatever grass is eaten by one animal is not available for another. It is also non-excludable since one herder cannot exclude others from using it. This story is used by environmentalists to illustrate the overuse of a resource when there are no restrictions on its use.

The tragedy of the commons has been challenged by the ideas of Elinor Ostrom, that we will study below under the topic *Collective self-governance* as well as in Theory of knowledge 5.1.

Sustainability and common pool resources

As we know from Chapter 1, *sustainability* in connection with the environment refers to the use of resources in ways that do not result in fewer or lower-quality resources for future generations. Sustainable production is production that uses resources in a sustainable way, in other words by not degrading or depleting them. By contrast, **unsustainable production** refers to production that uses resources unsustainably, depleting or degrading them.

Sustainable production: maximum sustainable yield of common pool resources (Supplementary material)

A simple example shown in Figure 5.1(a) illustrates the meaning of sustainable and unsustainable production and resource use. Fish in the open seas are a common pool resource that anyone has access to without payment. The horizontal axis measures the number of fishing boats, and the vertical axis measures the quantity of fish caught in tonnes. The first, second and third boats each catch 4 tonnes; therefore, in this range of 'constant average yield' (yield refers to the amount of output), the three boats together catch 12 tonnes, or 4 tonnes each on average.

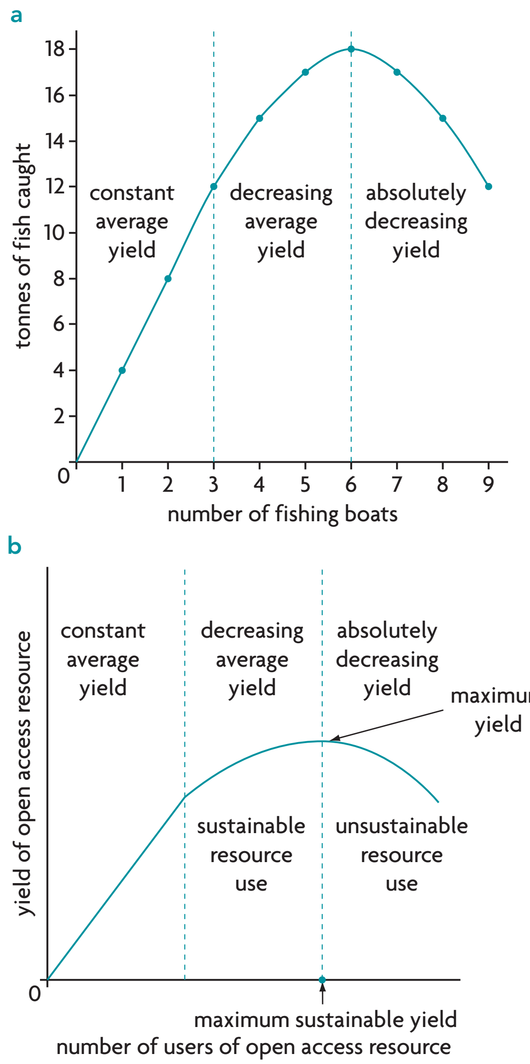


Figure 5.1: Illustrating sustainable and unsustainable resource use

When a fourth boat goes out to sea, it brings back only 3 tonnes of fish; this translates into a smaller quantity of fish caught by each boat on average. The four boats together have caught 15 tonnes, or an average of 3.75 tonnes ($= \frac{15}{4}$) instead of 4 tonnes.

When the fifth boat is added, the five boats catch 17 tonnes, and the average catch falls further to 3.4 tonnes ($= \frac{17}{5}$). With the sixth boat, the total is only 19 tonnes or 3.2 tonnes for each boat on average. This is the range of 'decreasing average yield', meaning that each boat that goes out brings back a smaller amount of fish than the previous one.

What happens if a seventh boat goes out? The total amount of fish caught by the seven boats together (17 tonnes) is *less* than what was caught by 6 boats (18 tonnes). As the graph indicates, in this range of 'absolutely decreasing yield', as more and more boats go fishing, the total amount of fish they bring back becomes less and less.

This example illustrates that the fish were plentiful for the first three boats, but with the addition of the fourth, fishing became more difficult because it began to put pressure on the supply of fish in the ocean. As the supply of fish was more and more depleted, it became increasingly difficult to catch fish, so the average quantity of fish brought back fell with the addition of each boat. Finally, with the addition of the seventh boat, the fish supply was *overused*; the fish population was no longer able to reproduce itself, and therefore the quantity of fish in the ocean began to drop

Figure 5.1(b) shows that the point of maximum yield of a common pool resource is the resource's *maximum sustainable yield. This is the maximum use that can be made of the resource that is also sustainable, in that the resource can reproduce itself.* All points to the left of the maximum sustainable yield indicate sustainable levels of use; points to the right indicate unsustainable use, meaning that the resource is being depleted or degraded. The further to the right, the greater the resource depletion or degradation. In the real world, many common pool resources are used unsustainably, i.e. to the right of their maximum sustainable yield.

Note that while it is an easy matter to discuss the maximum sustainable yield of a resource in theoretical terms as we have done here, it is very difficult in practice to determine what this actually is for any resource.

Sustainable resource use means that resources are used at a rate that allows them to reproduce themselves, so that they do not become degraded or depleted.

A note on renewable and non-renewable resources, and sustainability

Non-renewable resources are those resources that do not last indefinitely, because they have a finite supply (they need tens of thousands or millions of years to reproduce themselves). Examples include metals, minerals and fossil fuels, such as oil, natural gas and coal. Many of these resources, with the exception of fossil fuels, do not get destroyed through their use, and so through effective recycling could be made to last indefinitely. By contrast, fossil fuels are destroyed when used, and moreover have devastating effects on the earth's atmosphere, the global climate and the ozone layer.

Renewable resources are those resources that can last indefinitely if they are managed properly (not overused), because they are reproduced over relatively short periods of time by natural processes. Examples include forests, wildlife, fish, biomass, water resources, geothermal power, soil fertility and biodiversity. The idea of sustainable resource use applies mainly to *renewable resources*, because given appropriate management, these resources can be made to last forever. On the other hand, through mismanagement or overuse, these resources become depleted and degraded, indicating unsustainability.

The idea of sustainable resource use does not apply to non-renewable resources, such as fossil fuels. If resources are non-renewable, they could be used sustainably only if they were not used at all. On the other hand, as we will see in the pages that follow the idea of sustainability is relevant to fossil fuels when referring to the negative externalities that are created by their use.

TEST YOUR UNDERSTANDING 5.1

1 Provide examples of common pool resources, making reference to their overuse.

2 a Define common pool resources using the concepts of rivalry and non-excludability.

b Explain how these characteristics pose a threat to the environment.

c Outline how these characteristics can be illustrated by the tragedy of the commons.

3 Outline the meaning of sustainable production.

4 a In discussions of common pool resources, there is an emphasis on their *overuse* rather than their *use*. Explain why.

b Explain why cutting down a small amount of forest over an extended period of time may be consistent with the concept of environmental sustainability.

5.2 Market failure and externalities: diverging private and social benefits and costs

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain market failure as the failure of the market to achieve allocative efficiency (AO2)

CONTINUED

- explain that socially optimum output is produced when there is allocative efficiency; this occurs when MSC=MSB and social (community) surplus is maximum (AO2)
- draw a diagram illustrating allocative efficiency and socially optimum output (AO4)
- explain the meaning of externalities as an example of market failure (AO2)

Introducing market failure: the failure of the market to achieve allocative efficiency

Our discussion in previous chapters has shown that a free, competitive market economy gives rise to a number of desirable outcomes. In Chapter 2 we learned that in a free competitive market, when the price of a good adjusts to make quantity demanded equal to quantity supplied, the equilibrium quantity reflects the 'best' or optimal allocation of resources to the production of that good. This condition is known as allocative efficiency, achieved when marginal benefit equals marginal cost ($MB = MC$), or when social surplus is maximum.

However, the achievement of these outcomes depends on very strict and unrealistic conditions that are practically never met in the real world. Therefore, in reality, the free market most often fails to achieve these highly desirable results. The study of market failure focuses on one particular failing: the free market's inability to realise allocative efficiency in a variety of circumstances.

Market failure does not necessarily lessen the market's significance as a mechanism that can advance the well-being of societies; instead, it suggests that for markets to realise their potential, they must be supported by appropriate government policies. Allocative efficiency is a concept used by economists to identify real-world situations that differ from the ideal of a perfect allocation of resources. Once these are identified, it is possible to design government policies aimed at reducing the extent of the inefficiencies.

Market failure refers to the failure of the market to allocate resources efficiently. Market failure results in **allocative inefficiency**, where too much or too little of goods or services are produced and consumed from the point of view of what is socially most desirable. Overprovision of a good means too many resources are allocated to its production (overallocation); underprovision means that too few resources are allocated to its production (underallocation).

TEST YOUR UNDERSTANDING 5.2

1. Using a diagram, and the concepts of consumer and producer surplus, and marginal benefits and marginal costs, explain the meaning of allocative efficiency.
2. Explain, in a general way, the meaning of market failure.

The meaning of externalities: diverging private and social benefits and costs

Understanding externalities

When a consumer buys and consumes a good, she or he derives some benefits. When a firm produces and sells a good, it incurs costs. Sometimes the benefits or costs spill over onto other consumers or producers or society as a whole, who have nothing to do with consuming or producing the good. When this happens, there is an externality.

An **externality** occurs when the actions of consumers or producers give rise to negative or positive side-effects on other people who are not part of these actions, and whose interests are not taken into consideration.

The other people feeling the effects of an externality are often referred to as 'third parties'. If the side-effects on third parties involve benefits, there arises a **positive externality**, also known as external (or spillover) benefit; if they involve costs, in the form of negative side-effects, there arises a **negative externality**, also known as external (or spillover) costs.

Externalities can result either from consumption activities (consumption externalities) or from production activities (production externalities).

Marginal private benefits and costs, and marginal social benefits and costs

To fully understand externalities, let's return to the demand and supply curves we studied in Chapter 2. As we know, the demand curve is also a 'marginal benefit curve' where marginal benefit is the benefit received by consumers for consuming one more unit of the good (see Figure 2.17, Chapter 2). Since the benefits derived from consuming the good go to private individuals, who are the consumers buying the good, the demand curve represents marginal *private* benefits, shown as *MPB* in Figure 5.2.

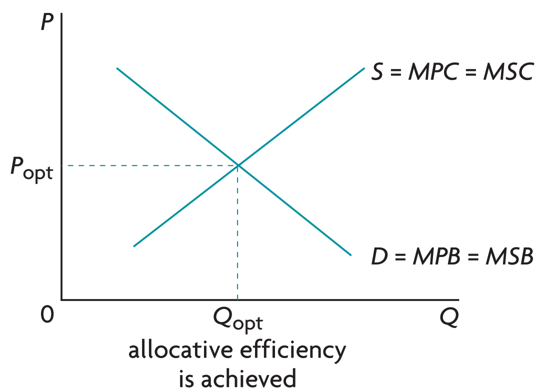


Figure 5.2: Demand, supply and allocative efficiency with no externalities

The standard supply curve reflects firms' costs of production, specifically marginal costs (see Figure 2.17). Marginal cost is the cost to producers of producing one more unit of the good. The supply curve therefore represents marginal *private* costs, appearing as *MPC* in Figure 5.2.

Now, if there are no externalities, so the actions of buyers and sellers do not produce side-effects on third parties, the marginal private benefit (*D*) curve and marginal private cost (*S*) curve determine an equilibrium price and quantity that reflect a *social optimum*, where there is allocative efficiency. In Figure 5.2, these are P_{opt} and Q_{opt}. A social optimum refers to a 'best' situation from the point of view of allocative efficiency. Q_{opt} is known as **socially optimum output**.

If, however, there is an externality, additional benefits or additional costs affecting third parties arise, and the full benefits or full costs to society differ from the private ones. These involve *marginal social benefits* (*MSB*) that differ from marginal private benefits; or *marginal social costs* (*MSC*) that differ from the marginal private costs.

When this occurs, the equilibrium price and quantity determined by the intersection of the demand (*MPB*) curve and supply (*MPC*) curve is no longer a social optimum, because *allocative inefficiency* is introduced by social benefits or costs that differ from private ones.

In a diagram, social benefits appear as a marginal social benefit curve, *MSB*, representing the full benefits to society from the consumption of a good, and social costs as a marginal social cost curve, *MSC*, representing the full costs to society of producing the good. When *MSB* and *MSC* are equal to each other, there is a social optimum in which allocative efficiency is realised.

Figure 5.2 shows the case where there are no external benefits or external costs (no externalities). Therefore $D = MPB = MSB$, and $S = MPC = MSC$.

> **marginal private costs** (*MPC*) refer to costs to producers of producing one more unit of a good.
>
> **marginal social costs** (*MSC*) refer to costs to society of producing one more unit of a good.
>
> **marginal private benefits** (*MPB*) refer to benefits to consumers from consuming one more unit of a good.
>
> **marginal social benefits** (*MSB*) refer to benefits to society from consuming one more unit of a good.

> When $MSC = MSB$, there is allocative efficiency and *socially optimum output* is produced. When there is no externality, the competitive free market leads to an outcome where $MPC = MSC = MPB = MSB$, as in Figure 5.2, indicating allocative efficiency. An externality creates a divergence between *MPC* and *MSC* or between *MPB* and *MSB*. When there is an externality, the free market leads to an outcome where $MPB = MPC$, but where *MSB* is not equal to *MSC*, indicating allocative inefficiency. Either too much or too little is produced relative to the social optimum.

You should note that the condition $MSC = MSB$ is the same as $MC = MB$ when there are no externalities. In Chapter 4 we referred to $MC = MB$ as the condition for allocative efficiency because we were considering markets with no market failures.

We will examine four types of externalities:

- negative production externalities
- negative consumption externalities
- positive production externalities
- positive consumption externalities.

The first two of these, negative externalities of production and consumption, will be examined in the present chapter. The last two, positive externalities of production and consumption, will be examined in Chapter 6.

The present chapter will also examine in more detail the topic of *common pool resources*, introduced above, which are very closely related to negative externalities.

TEST YOUR UNDERSTANDING 5.3

1. a. Outline the meaning of an externality.
   b. Use the concept of allocative efficiency to explain how externalities relate to market failure.
2. Explain the difference:
   a. between marginal private benefit and marginal social benefit, and
   b. between marginal private cost and marginal social cost.
3. State what condition of perfect markets is violated, leading to allocative inefficiency, when there is an externality.

5.3 Negative production externalities

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain negative externalities of production and the resulting welfare loss (AO2)
- draw a diagram illustrating negative externalities of production and welfare loss (AO4)
- calculate welfare loss that arises from negative externalities of production (HL only) (AO4)
- explain that negative production externalities can be used to illustrate overuse of common pool resources (AO2)
- explain government intervention to correct negative externalities of production and prevent overuse of common pool resources including: (AO2)
    - indirect (Pigouvian) taxes
    - carbon taxes
    - tradable permits
    - legislation and regulation
    - collective self-governance
    - education-awareness creation
    - international agreements
- draw diagrams to illustrate the above government responses (AO4)
- discuss strengths and limitations of the above government policies with respect to: (AO3)
    - difficulties in measurement of externalities
    - degree of effectiveness
    - consequences for stakeholders

Explaining and illustrating negative production externalities

Negative production externalities refer to external costs created by producers. The problem of environmental pollution, created as a side-effect of production activities, is very commonly analysed as a negative production externality.

Consider a cement factory that emits smoke into the air and disposes its waste by dumping it into the ocean. There is a production externality, because over and above the firm's private costs of production, there are additional costs that spill over onto society due to the polluted air and ocean, with negative consequences for the local inhabitants, swimmers, sea life, the fishing industry and the marine ecosystem. This is shown in Figure 5.3, where the supply curve, $S = MPC$, reflects the firm's private costs of production, and the marginal social cost curve given by MSC represents the full cost to society of producing cement. For each level of output, Q, the social costs of producing cement given by MSC are greater than the firm's private costs. The vertical difference between MSC and MPC represents the external costs. Since the externality involves only production (the supply curve), the demand curve represents both marginal private benefits and marginal social benefits.

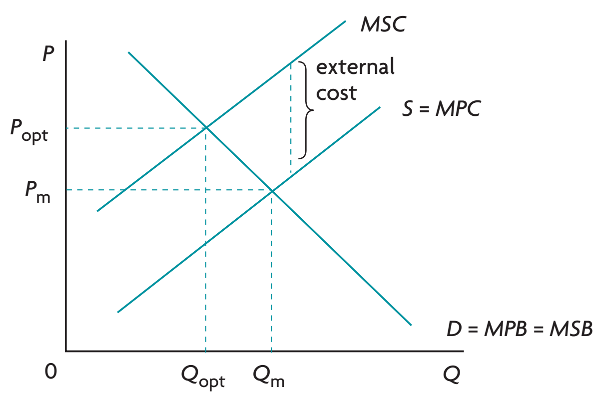


Figure 5.3: Negative production externality

Figure 5.3 illustrates a general point that you should keep in mind whenever you examine (or draw) an externality diagram: the free market outcome is determined by the intersection of MPB and MPC, resulting in quantity Q_m and price P_m. The socially optimum (or 'best') outcome is given by the intersection of MSB with MSC, which determines quantity Q_{opt} and price P_{opt}.

We can draw an important conclusion from the negative externality in Figure 5.3:

> When there is a negative production externality, the free market overallocates resources to the production of the good and too much of it is produced relative to the social optimum. This is shown by $Q_m > Q_{opt}$ and $MSC > MSB$ at the point of production, Q_m, in Figure 5.3.

The welfare loss of negative production externalities

Welfare loss

Whenever there is an externality, there is a welfare loss, involving a reduction in social benefits, due to the misallocation of resources.

In Figure 5.4(a), the shaded area represents the welfare loss arising from the negative production externality. For all units of output greater than Q_{opt}, $MSC > MSB$, meaning that society would be better off if less were produced. The welfare loss is equal to the difference between MSC and MSB for the amount of output that is overproduced ($Q_m - Q_{opt}$). It is a loss of social benefits due to overproduction of the good caused by the externality. If the externality were corrected, so that the economy reaches the social optimum, the loss of benefits would disappear. It may be useful to note that the point of the welfare loss triangle always lies at the Q_{opt} quantity of output.

Calculating welfare loss (HL only)

Figure 5.4(b) is similar to 5.4(a) except that it includes figures that allow us to calculate welfare loss. To find the area of the shaded triangle, we take the height times the width of the triangle and divide it by 2. Note that the height of the triangle is equal to the external cost per unit, or $MSC - MPC$, and the width is equal to the amount of overproduction by the market or $Q_m - Q_{opt}$:

$$\text{Welfare loss} = \frac{(6-4)\times(100-70)}{2} = \frac{2\times 30}{2} = \$30$$

Welfare loss in relation to consumer and producer surplus (Supplementary material)

We can use the concepts of consumer and producer surplus to understand the welfare loss due to the externality. In Figure 5.4(c), in market equilibrium,

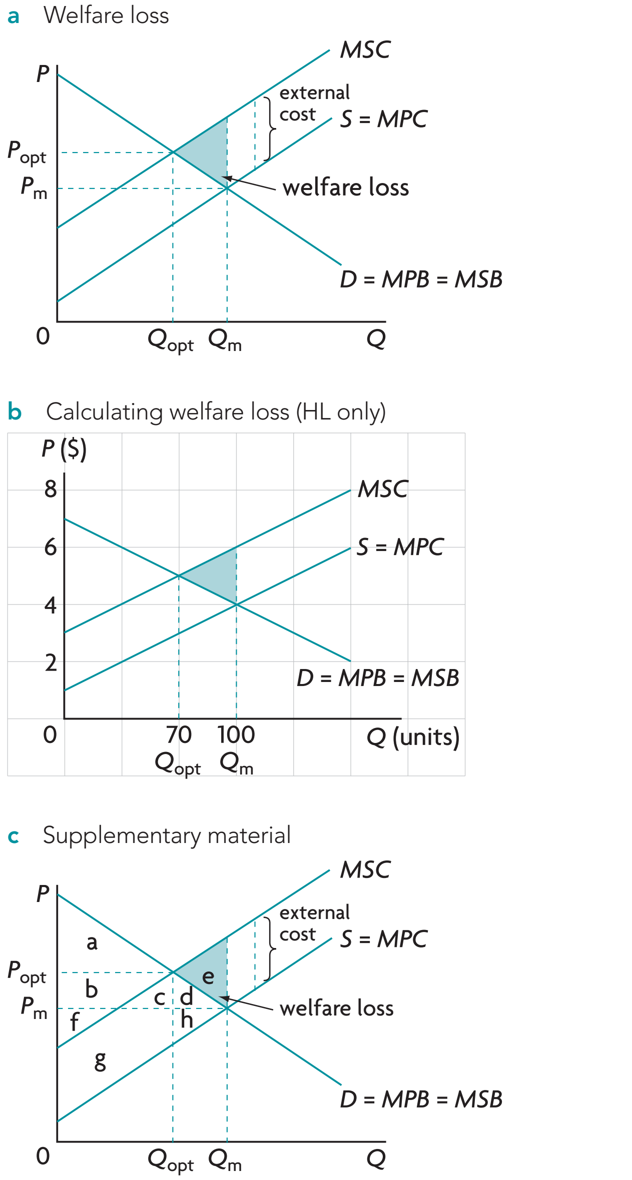


Figure 5.4: Welfare loss in a negative production externality

consumer surplus is equal to areas a + b + c + d, while producer surplus is equal to areas f + g + h. The value of the external cost is the difference between the MSC and MPC curves up to Q_m (the quantity produced by the market), and is therefore equal to c + d + e + g + h. The total social benefits in market equilibrium are equal to consumer surplus plus producer surplus minus the external cost:

$$(a + b + c + d) + (f + g + h) - (c + d + e + g + h)$$
$$= a + b + f - e$$

At the social optimum, or at Q_{opt} and P_{opt}, consumer surplus is equal to area a, and producer surplus is equal to area b + f. The external cost is now equal to zero. Therefore, the total social benefits are equal to consumer surplus plus producer surplus:

a + b + f

Comparing total social benefits at the market equilibrium and at the social optimum, we find that they are smaller at the market equilibrium by the area e. This is the welfare loss.

Negative production externalities and overuse of common pool resources

The concept of negative externalities that we studied above can be used to illustrate the problem of overuse of common pool resources and its effects on sustainability. For example, in the negative production externality diagram of Figure 5.3 the difference between the *MPC* and *MSC* curves can be interpreted as the external cost arising from the cement factory's overuse of clean air, water and sea life on account of its dependence on fossil fuels; it can also be interpreted as the cost to society of causing global warming (destroying the stability of the global climate, which is also a common pool resource). The burning of fossil fuels creates external costs in terms of overuse of common pool resources.

If it were possible to make the cement factory pay for the overuse of these resources, the producer would not necessarily stop its polluting activities entirely, and would not stop *using* common pool resources. However, it would stop *overusing* them, thus leading to a sustainable use of common pool resources.

Figure 5.3 can be used to illustrate the overuse of many common pool resources as a negative production externality. For example, if the S = *MPC* curve represents the private costs of a fishing firm that fishes in the open seas, the external costs would be depletion of the stock of fish, and environmental damage due to disruption of the marine ecosystem, or the common pool resources that the fishing firm has overused but not paid for. The costs to society of the firm's fishing activities are given by *MSC*, which are the private costs plus the external costs.

TEST YOUR UNDERSTANDING 5.4

1. a. Using a diagram, show how marginal private costs and marginal social costs differ when there is a negative production externality.
   b. Explain the difference between the equilibrium quantity determined by the market and the quantity that is optimal from the point of view of society's preferences.
   c. Describe the problem with the allocation of resources achieved by the market when there is a negative production externality.
   d. Show the welfare loss created by the negative production externality in your diagram, and explain what this means.
2. a. Provide examples of negative production externalities.
   b. Using diagrams and examples, explain how negative production externalities can be used to analyse the overuse of common pool resources.

Policies to correct negative production externalities and prevent overuse of common pool resources and their evaluation

We will examine and evaluate a variety of policies to deal with negative production externalities and overuse of common pool resources.

Market-based policies I: indirect (Pigouvian) taxes

An important group of policies that can be pursued by governments rely on the market to correct negative production externalities and promote sustainable use of common pool resources. Market-based policies work by changing the incentives faced by firms.

In one such approach the government could impose an indirect tax on the firm per unit of output produced.

This is known as a **Pigouvian tax** (or **Pigovian tax**), after the English economist Arthur Cecil Pigou who was the first to propose the idea of imposing a tax in order to correct a negative externality. In Figure 5.5(a), the tax results in an upward shift of the supply curve, from $S = MPC$ to MSC ($=MPC$ + tax). The optimal (or best) tax policy is to impose a tax that is exactly equal to the external cost, so the MPC curve shifts upward until it overlaps with MSC. The new, after-tax equilibrium is given by the intersection of MSC and the demand curve, $D = MPB = MSB$, *resulting in the lower, optimal quantity* of the good produced, Q_{opt}, and higher, optimal price, P_{opt}. Note that whereas the indirect taxes discussed in Chapter 4 *introduced allocative inefficiency*, indirect taxes in the present context are intended to *lead to allocative efficiency.*

Bearing in mind our discussion of indirect taxes in Chapter 4, you may note that P_{opt} is the price paid by consumers, or P_c, while the price received by producers is P_p, which is equal to P_c minus tax per unit.

Market-based policies II: carbon taxes

Carbon taxes are a kind of tax designed to deal with what is perhaps the single most pressing and complex threat to the global ecosystem: global warming, caused by emissions of greenhouse gases, the most important of which is carbon dioxide. When we speak of the contribution of greenhouse gases to global warming, we refer to those gases emitted by man-made processes, and specifically by the burning of fossil fuels (oil, coal and natural gas).

There are two key measures under discussion in the global community to deal with the problem of carbon dioxide emissions: carbon taxes and tradable permits (see below).

The **carbon tax** is a tax per unit of carbon emissions of fossil fuels. Fossil fuels do not all emit the same amounts of carbon when burned, therefore the carbon tax is calculated on the basis of how much carbon the fuel emits: *the more carbon emitted, the higher the tax*. This can be illustrated by the same diagram used to show the effects of a tax per unit of output, in Figure 5.5(a). Following the imposition of the tax, firms must pay the higher price to buy the fossil fuel. This appears in Figure 5.5(a) as the familiar upward shift in $S = MPC$ toward MSC because of the firm's higher costs of production, but this has further consequences. Since there are other substitute energy sources with lower carbon emissions (thus taxed at a lower rate), or that do not emit carbon (if they are not fossil fuels, thus not taxed at all), the increase in the price of the high-carbon fuel creates incentives for firms to switch to other, less polluting or non-polluting energy sources.

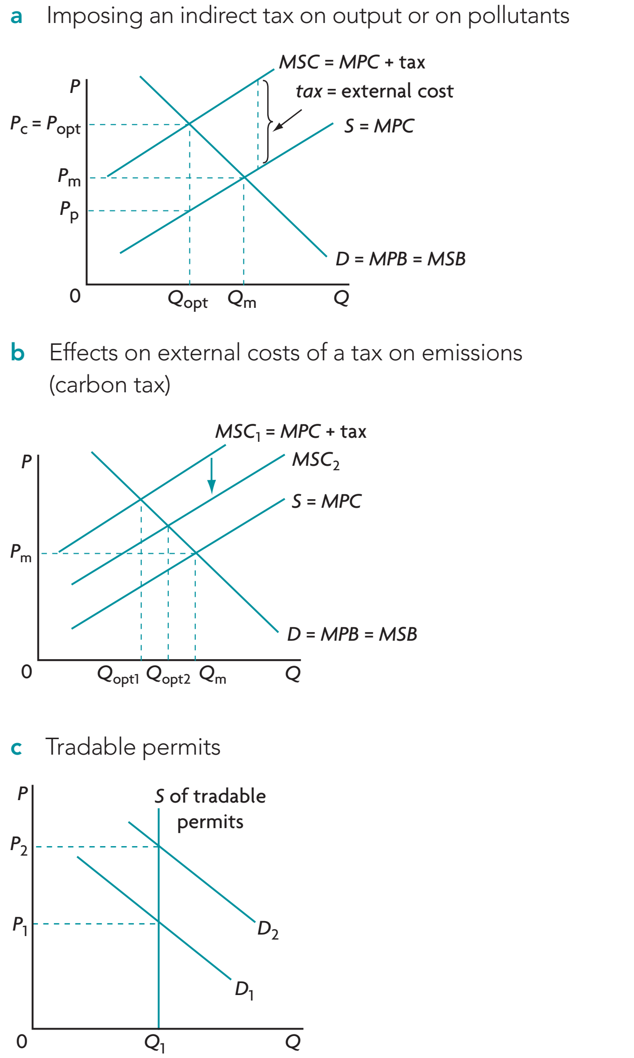


Figure 5.5: Market-based policies to correct negative production externalities and promote sustainable use of common pool resources

The result is that if the firm switches to alternative, less polluting resources, Q_{opt} *in Figure 5.3 will increase, because the external costs of producing the output will become smaller.* This can be seen in Figure 5.5(b), where the MSC curve shifts from MSC_1 to MSC_2, indicating that the external costs are lower due to the use of the less polluting resources. With the fall in external costs, the optimum quantity of output increases from Q_{opt1} to Q_{opt2}. (Note that this also involves a lower tax on pollutants, shown by the smaller distance between the demand curve and MSC_2.)

Carbon taxes are used in many countries as a method to reduce pollution, for example, Denmark, Finland, France, Ireland, Japan, Mexico, Poland, Sweden, as well as some states in Canada and the United States and more.

A tax on carbon (or on emissions generally) has the effect of creating incentives for producers to reduce the amount of pollution they create by purchasing less polluting resources (such as fossil fuels), and to switch to less polluting technologies (alternative energy resources). This reduces the size of the negative externality and increases the optimum quantity of output. A tax on the output of the polluter does not have this effect; it corrects the overallocation of resources to the good, reducing the quantity of output produced.

Market-based policies III: tradable permits

Tradable permits, also known as *cap and trade schemes*, are a policy involving *permits to pollute* issued to firms by a government or an international body. These permits to pollute can be traded (bought and sold) in a market. Consider a number of firms whose production pollutes the environment. The government grants each firm a particular number of permits (or rights) to produce a particular level of pollutants over a given time period. The permits to pollute can be bought and sold among interested firms, with the price of permits being determined by supply and demand. If a firm can produce its product by emitting a lower level of pollutants than the level set by its permits, it can sell its extra permits in the market. If a firm needs to emit more pollutants than the level set by its permits, it can buy more permits in the market.

Figure 5.5(c) shows a market for tradable pollution permits. The supply of permits is perfectly inelastic (i.e. the supply curve is vertical), as it is fixed at a particular level by the government (or an international authority if several countries are participating). For this policy to effectively reduce the level of pollution, the total pollution that is permitted based on the pollution permits must be *less than the amount of pollution created with no permits*. The fixed supply of permits is distributed to firms. The position of the demand-for-permits curve determines the equilibrium price. As an economy grows and the firms increase their output levels, the demand for permits is likely to increase, as shown by the rightward shift of the demand curve from D_1 to D_2. With supply fixed, the price of permits increases from P_1 to P_2.

Tradable permits are like taxes on emissions in that they provide incentives to producers to switch to less polluting resources for which it is not necessary to buy permits. If a firm finds a way to reduce its emissions, it can sell its permits thus adding to profits. Permits are therefore intended to reduce the quantity of pollutants emitted, thus reducing the size of the negative externality, and increasing the optimum quantity of output produced, by shifting the *MSC* curve upward toward *MPC*, as shown in Figure 5.5(b).

Tradable permit schemes may be set up within a country, such as Kazakhstan (2013), Switzerland (2008), New Zealand (2008); or within a groups of countries such as the European Union Emissions Trading System (EU ETS) (2005); or internationally such as the Paris Agreement (2016; see below *International agreements*). Tradable permits are hotly debated together with carbon taxes; see Real World Focus 5.1.

You may note in your reading about tradable permits that they are referred to as an Emissions Trading System (ETS).

Advantages of market-based policies

Economists usually prefer the market-based solutions to government regulations to deal with negative production externalities and overuse of common pool resources, as long as the particular externality permits the use of such policies. Both taxes and tradable permits have the effect of *internalising the externality*, meaning that the costs that were previously external are made internal, because they are now paid for by producers and consumers who are parties to the transaction.

In the case of taxes, *taxes on emissions such as the carbon tax are superior to taxes on output*. Taxes on output only provide incentives to producers to reduce the quantity of output produced with a given technology and given polluting resources, but not to reduce the amount of pollution they create or to switch to less polluting resources.

Taxes on pollutants emitted provide incentives to firms to economise on the use of polluting resources (such as fossil fuels) and use production methods that pollute less. Firms do not all face the same costs of reducing pollution; for some, the costs of reducing pollution are lower than for others, and these will be the ones most likely to cut their pollution emissions to avoid paying the tax. Firms that face the highest costs of reducing pollution will be the ones least likely to cut their pollutants, and so will pay the tax. The result is that *taxation leads to lower pollution levels at a lower overall cost to society* since the firms that will switch to clean forms of energy are the ones that can do it more cheaply.

Similarly, in the case of tradable permits, the system creates incentives for firms to cut back on their pollution if they can do so at relatively low cost. If it is a relatively low-cost procedure for a firm to reduce its pollutant emissions, it will be in its interests to do so and sell excess permits. Firms that can only reduce pollution at high cost will be forced to buy additional permits. Therefore, both taxes and tradable permits are methods to reduce pollution more efficiently (at a lower cost).

Disadvantages of market-based policies

Whereas taxes and tradable permits are simple in theory, in practice they are faced with numerous technical difficulties. While it is known with a reasonable degree of certainty that man-made greenhouse gases cause global warming, there is tremendous uncertainty in calculating the precise contribution of each of these to increases in global temperatures. This gives rise to difficulties in designing effective carbon taxes and tradable permit schemes.

Taxes

Taxes face serious practical difficulties that involve designing a tax equal in value to the amount of the pollution. An effective tax policy requires answers to the following questions:

- **What production methods produce pollutants?** Different production methods create different pollutants. It is necessary to identify what methods produce which pollutants, which is technically very difficult.
- **Which pollutants are harmful?** It is necessary to identify the harmful pollutants, which is also technically difficult, and there is much controversy among scientists over the extent of harm done by each type of pollutant.
- **What is the value of the harm?** It is then necessary to attach a monetary value to the harm: how much is the harm done by each pollutant worth? This raises questions that have no easy answers: who or what is harmed; how is the value of harm to be measured?
- **What is the appropriate amount of tax?** It is necessary to determine the size of the tax to make it equal to the value of the harm.
- **How will consumers be affected?** Indirect taxes are *regressive* (as you will learn in in Chapter 12) meaning that lower income people have to pay a higher proportion of their income in tax than higher income people, which is considered inequitable (unfair).

A serious problem with carbon taxes is that they are usually set too low to make a significant impact. According to the OECD Secretary General:

> *The gulf between today's carbon prices and the actual cost of emissions to our planet is unacceptable. Pricing carbon correctly is a concrete and cost-effective way to slow climate change. We are wasting an opportunity to steer our economies along a low-carbon growth path and losing precious time with every day that passes.*[2]

The reason is that it is politically difficult to impose carbon taxes that are high enough to make the necessary difference.

Aside from the technical difficulties, there is also a risk that even if taxes are imposed some polluting firms may not lower their pollution levels, continuing to pollute even though they pay a tax.

Tradable permits

Tradable permits face the technical limitations regarding production methods and pollutants noted above for taxes. In addition, tradable permits require the government (or international body) to set a maximum acceptable level for each type of pollutant, called a 'cap'. This task demands having technical information on quantities of each pollutant that are acceptable from an environmental point of view, which is often not available. If the maximum level is set too high, it will not have the desired effect on cutting pollution levels. If it is set too low, the permits become very costly, causing hardship for firms that need to buy them. To date, tradable permits have been developed for just a few pollutants (CO_2, SO_2).

In addition, a method must be found to distribute permits to polluting firms in a fair way. Issues of political favouritism may come into play, as governments give preferential treatment to their 'friends' and supporters.

In practice, the most that can be hoped for is a shift of the *MPC* curve toward the *MSC* curve, as well some reduction in the size of the externality, but it is unlikely that these policies can achieve the optimal results.

[2] OECD, Few countries are pricing carbon high enough to meet climate target
http://www.oecd.org/tax/few-countries-are-pricing-carbon-high-enough-to-meet-climate-targets.htm

REAL WORLD FOCUS 5.1

Carbon taxes versus tradable permits: how to best limit carbon dioxide emissions

There is a growing momentum around the world to implement carbon pricing in some form. As of 2019 over 60 countries, states or cities had implemented some form of carbon pricing, either through carbon taxes or through tradable permits. A survey of people in five countries (Australia, India, South Africa, the United Kingdom and the United States) found that between 60% (in the United States) and 80% (in India) supported carbon taxes provided these were on a global scale and the revenues were returned to the people or spent on climate projects.

Figure 5.6: Klitten, Germany. Lignite-fired power station

However, carbon pricing remains unpopular among both consumers and businesses. Businesses are often opposed because it raises the costs of production. Others say that it is unfair in the event that they trade with countries that do not have carbon pricing. In Australia, a carbon tax was repealed in 2014 on the grounds that it was destroying jobs. It is often opposed by consumers because it has the effect of raising prices. It has a stronger effect on the poor because it raises household energy prices.[4]

As market-based methods to reduce emissions, both carbon taxes and tradable permits provide incentives to firms to switch to less polluting forms of energy. However, as we have seen they differ in how they attempt to do this. Carbon taxes fix the price of the pollutant in the form of a tax on carbon and allow the quantity of carbon emitted to vary, depending on how firms respond to the tax; cap and trade schemes fix the quantity of the permissible pollutant, and allow its price to vary, depending on supply and demand.

Carbon taxes versus tradable permits: an evaluation

- **Carbon taxes make energy prices more predictable.** Fossil fuel prices in global markets fluctuate according to demand and supply. Under tradable permits, the price of fossil fuels might fluctuate even more due to fluctuations in the price of carbon permits. Price predictability is important for businesses that need to plan their costs ahead of time.
- **Carbon taxes are easier to design and implement.** Tradable permits are difficult to design and implement as they involve complicated decisions such as setting the cap at the right level and distributing the permits among all interested users.
- **Carbon taxes can be applied to all users of fossil fuels.** Tradable permit schemes often target one particular industry, or small group of industries. Carbon taxes can be applied to all users of fossil fuels, including all producers and consumers.
- **Carbon taxes do not offer opportunities for manipulation by governments and interest groups.** Politicians often prefer tradable permit schemes to carbon taxes, and it is believed that this may be because it is easy to manipulate the distribution of permits for the benefit of preferred groups and supporters, without affecting the impacts on the environment (because of the cap).

[3] https://www.sciencedaily.com/releases/2019/01/190117092556.htm

[4] https://www.bloomberg.com/news/articles/2018-10-14/why-pricing-carbon-is-still-more-theory-than-reality-quicktake

CONTINUED

- **Carbon taxes do not require as much monitoring for enforcement.** Tradable permit schemes require monitoring of emissions, otherwise firms may try to cheat by emitting more pollutants than they are permitted. Carbon taxes are easier to monitor as they only involve payment of a tax depending on the type and quantity of fossil fuels purchased.
- **Tradable permit schemes face political pressures to set the cap too high.** If the cap on pollutants is set too high, it would have a very limited or no impact on reducing carbon emissions.

There are also some arguments against carbon taxes and in favour of tradable permit schemes:

- **Carbon taxes face political pressures to be set too low.** Governments may be unwilling to set carbon taxes high enough for these to provide the necessary incentives for users to switch to less polluting energy sources.
- **Carbon taxes cannot target a particular level of carbon reduction.** Since carbon taxes cannot fix (or cap) the permissible level of carbon emissions, they lead to uncertain carbon-reducing outcomes.
- **Carbon taxes are regressive.** A regressive tax is one where the tax as a fraction of income is higher for low-income earners than it is for higher-income earners, and go against the principle of equity (see Chapter 12). A carbon tax on a firm is an indirect tax that is paid partly by producers and partly by consumers. Therefore, consumers would also be affected, and lower-income consumers would be affected proportionately more.

Applying your skills

1. Compare and contrast some key issues surrounding the debate on carbon taxes versus cap and trade schemes.
2. The World Bank regularly updates the following site with information on carbon pricing (carbon taxes or tradable permits) that countries around the world are either implementing or planning to implement. Select one or more countries of your choice and investigate the carbon pricing that has been selected. Research the experience of your country or countries with respect to (a) political acceptability, (b) effectiveness with respect to reducing carbon emissions, (c) any future plans for tackling the carbon emissions problem.

   Source: *https://carbonpricingdashboard.worldbank.org*

Government legislation and regulation

Government legislation and regulations rely on the 'command' approach, where the government uses its authority to enact legislation and regulations in the public's interest (see Chapter 1 for a discussion of command decision-making).

Legislation and regulations intended to reduce the effects of production externalities and limit environmental damage typically involve emissions standards, quotas, licences, permits or outright restrictions. Examples include:

- restrictions on emissions of pollutants from factories and industrial production by setting a maximum level of pollutants permitted
- requirements for steel mills and electricity generating plants to install smokestack scrubbers to reduce emissions
- banning the use of harmful substances (e.g. asbestos, a group of minerals that are very toxic that is banned in many countries)
- issuing licences or permits for particular activities (such as hunting)

- prohibiting construction (such as housing) or industry or agriculture in protected areas
- restrictions on the quantity of logging
- restrictions in the form of quotas for fishing (maximum permissible quantity of fish that can be caught) or in the form of the size of shipping fleets, or total bans for specific areas or specific times of the year
- establishment of protected areas for the protection of biodiversity and endangered ecosystems.

The impact is to lower the quantity of the good produced and bring it closer to Q_{opt} in Figure 5.7 by shifting the *MPC* curve upward towards the *MSC* curve. Pollutant and output restrictions achieve this by forcing the firm to produce less. Requirements to install technologies reducing emissions achieve this by imposing higher costs of production due to the purchase of the non-polluting technologies. Ideally, the higher costs of production would be equal to the value of the negative externality. The government's policy objective is to make the *MPC* curve shift upwards until it coincides with the *MSC* curve, in which case Q_{opt} is produced, price increases from P_m to P_{opt}, and the problem of overallocation of resources to the production of the good is corrected. If polluting firms do not comply with the regulations, they would have to pay fines.

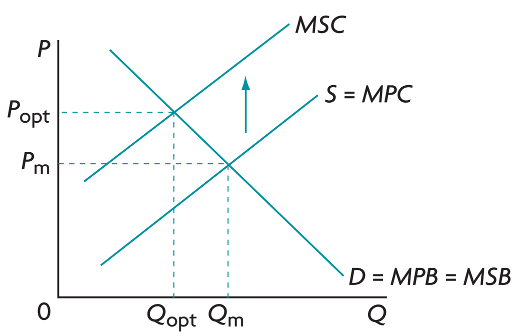


Figure 5.7: Government regulations to correct negative production externalities and promote sustainable use of common pool resources

Advantages

Legislation and regulations, including restrictions such as in the examples above, have the advantage that they are simple to put into effect and oversee. They are easier to implement compared to market-based polices and avoid the technical difficulties that arise in the use of market-based solutions. They can also be quite effective. For example, banning harmful substances, or prohibiting hunting in certain areas, or restricting the quantity of logging (chopping down trees) may be the most effective way to deal with certain problems. Moreover, regulations force firms to comply and reduce their harmful activities (which market-based policies may not always do). For these reasons, regulations are far more commonly used as a method to limit negative externalities of pollution in countries around the world.

Disadvantages

However, they also face limitations. In the case of emissions of pollutants, they do not offer incentives to reduce emissions by using less polluting resources, to increase energy efficiency and to switch to alternative fuels. They cannot distinguish between firms that have lower or higher costs of reducing pollution, which would limit the overall cost of reducing pollution (explained above). The result is that pollution is reduced at a higher overall cost.

In addition, although they can be implemented more easily, they suffer from similar limitations as the market-based policies (lack of sufficient technical information on types and amounts of pollutants emitted), and so can in most cases be only partially effective in reducing the pollution created. Finally, there are costs of monitoring and supervision to detect possible violations, leading to opportunity costs, and there may be problems with enforcement. Therefore, such measures can only attempt to partially correct the problem.

Overall, the effectiveness of legislation must be assessed in relation to the particular use for which it is intended, as it can be more effective in some situations than in others.

> Correction of negative production externalities by market-based approaches or government legislation and regulation usually involve shifting the *MPC* curve upward toward the *MSC* curve through a variety of policies. For allocative efficiency to be achieved, the quantity of the good produced and consumed must fall to Q_{opt} as price increases to P_{opt}.[5]

[5] See 'Quantitative techniques' chapter in the digital coursebook for an explanation of the equivalence of upward and leftward shifts of the supply curve.

Collective self-governance: the contribution of Elinor Ostrom

The concept of **collective self-governance** refers to a solution to the use of common pool resources where the users take control of the resources and use them in a sustainable way. This concept runs counter to the idea of the *tragedy of the commons*, discussed above, where each user makes use of the resource at the expense of others with the result to ultimately degrade and deplete it.

Collective self-governance, as the term suggests, is not a policy that is imposed by the government. It is rather an approach to manage resources undertaken by communities of resource users by themselves, because they realize that it is in their own best interests to work collectively for the preservation of resources that are vital to their livelihood.

Collective self-governance as an approach to common pool resources was made famous by Elinor Ostrom, an American political scientist who became the first woman to receive the Nobel Prize in Economics (2009) for her work on the management of common pool resources. Ostrom spent decades studying how communities organise themselves to manage common resources such as fish stocks, woods, lakes, pastures and groundwater basins. She discovered that communities often develop complex mechanisms on how to make decisions and how to enforce rules, resulting in outcomes that differ widely from those predicted by standard economic theory. In an interview in 2010, she said:

> *If you are in a fishery or have a pasture and you know your family's long-term benefit is that you don't destroy it, and if you can talk with the other people who use that resource, then you may well figure out rules that fit that local setting and organise to enforce them. But if the community doesn't have a good way of communicating with each other or the costs of self-organisation are too high, they won't organise, and there will be failures.*[6]

Ostrom and her colleagues performed hundreds of experiments, which show that when people are users of a common pool resource, they are often able to find solutions on how to manage it sustainably, without overusing and destroying it, *provided they have good methods of communication between them.*[7] They make rules about how each user should behave and contribute to the management of the resource, and rules of enforcement and sanctions against those who violate the rules. Many studies by Ostrom and other scholars of real-life situations, involving the management of common pool resources, confirm the results of these experiments.

However, one requirement for sustainable resource management is that there must be boundaries of an area (such as a pasture, a wood or a lake, and so on) corresponding to the area that the resource users are managing. This raises the question whether Ostrom's approach is suitable for dealing with global problems such as the oceans and climate change. She herself notes the following:

> *I really despair over the oceans . . . It's so tempting to go along the coast and scoop up all the fish you can and then move on. With very big boats you can do that. I think we could move toward solving that problem, but right now there are not many instrumentalities for doing that.*

Advantages

Elinor Ostrom's work has shown that people do not always act in the self-interested, narrow manner presupposed by the tragedy of the commons. Instead, by working cooperatively, people can find solutions to the problem of overuse of common pool resources without top-down solutions imposed by governments. These cooperative solutions can be achieved in the absence of private ownership of resources (emphasised by supporters of free market economies where this is an essential characteristic) as well as in the absence of government-owned property (emphasised by supporters of command or planned economies; see Chapter 1 for these distinctions). However, it is important to have a legal system of *land rights* in place (see Chapters 19 and 20).

Disadvantages

For people to be able to manage common pool resources on their own there are two important conditions that must be satisfied, they must be able to communicate with each other in order to create rules for the use of the common pool resources, and there must be a boundary for the resource. As Ostrom herself admits this is difficult for resources such as the oceans.

Education and awareness creation

Education of the public and provision of information regarding the polluting activities of firms (or other activities with negative external effects) often makes consumers turn away from the products, with negative effects on the firms' sales. As a result the firms are forced to take consumers'

[6] Interview with Elinor Ostrom, by Fran Korten, Yesmagazine.org, 27 February 2010.
[7] The experiments are carried out at the Workshop in Political Theory and Policy Analysis at Indiana University where Ostrom was a professor.

opinions into consideration and change their production methods in order to reduce the externalities.

Advantages

The advantage here is that firms are very much influenced by the opinions of their customers and want to keep them happy, otherwise they will suffer drops in sales.

Disadvantages

The difficulty with this approach is that it can only make a small difference in terms of solving the problem of production externalities and sustainability. For example, if there is information about a firm creating significant environmental damage in a localised area (such as causing an oil spill that affects the livelihood and health of the local inhabitants), consumers may become concerned and boycott the firm for a while. However problems of a more general and broader nature, such as the use of fossil fuels that causes climate change plus a host of other environmental problems, require solutions on a far broader scale. In fact carbon taxes and tradable permits (discussed above) are far more effective in tackling problems of this kind.

International agreements

Policies are made mainly by national governments. However, negative production externalities and the overuse of common pool resources very often have international repercussions, in which case co-operation among governments and international agreements are crucially important to control and prevent negative consequences on certain resources, such as the global climate and the ozone layer. In addition, co-operation among governments is very important for the development and diffusion of new technologies intended to deal with global environmental issues. Co-operation between governments may be global or regional.

For example, the ozone layer has suffered ozone depletion, leading to reduced protection against the sun's ultraviolet radiation. This resulted from human activities involving the production of nitrogen oxides and chlorofluorocarbons (CFCs). The ozone layer is a common pool resource. No one owns it, and no one can claim damages for its destruction. The responsibility for its destruction lies with polluting activities within virtually every country, and the consequences of its destruction are felt globally. The same considerations apply to the global climate.

One of most successful examples of international collaboration for the environment is the Montreal Protocol, signed in 1987 and coming into effect in 1989, intended to phase out substances that have caused depletion of the ozone layer. By 2009, all member states of the United Nations had ratified the agreement, and significant progress has been made in the area of phasing out ozone-depleting substances.

Another successful example of a regional collaborative arrangement is the European Union's tradable permits scheme for carbon, known as the European Union Emissions Trading System (EU ETS), which was initiated in January 2005. The scheme covers the sectors of power and heat generation, oil refineries, metals, pulp and paper, and energy intensive industry. In this system, one permit, or EU Allowance (EUA) permits the holder to release one tonne of carbon dioxide. Each emitter of carbon is allocated EUAs, which are traded in a rapidly growing carbon market. The EU ETS is the cornerstone of the European Union's policy on climate change. According to a major study, while this has helped in reducing carbon emissions it has not had negative impacts on the economic performance of firms in terms of revenues, profits and employment.[8]

Another major, but less successful international agreement for the environment was the Kyoto Protocol of 2005–2012. Its objective was to make signatory countries commit themselves to reduce emissions of carbon dioxide and other greenhouse gases to slow down climate change. It also contained provisions for the development of a market of tradable emissions permits. The countries that signed were divided into developed and developing. Only the developed countries had emissions restrictions. The developing countries participated by investing in projects that were supposed to lower their emissions. The United States, which was the country with the highest greenhouse gas emissions, did not sign the Kyoto protocol on the grounds that it was facing unfair competition since developing countries did not face restrictions. China, the second highest greenhouse emitter, did not participate as it was not in the group of developed countries. Many environmental specialists argued that the agreed reductions in emissions were too small to have sufficient impact on the problem of global warming.

In 2016, the Paris Agreement came into effect, initially signed by 55 parties, which reached 197 by 2019. The purpose of the agreement is to strengthen international co-operation on climate change, based on the goal of limiting the global temperature increase to 1.5%. In addition it aims to increase the ability of countries to adapt to negative effects of climate change. It establishes binding agreements on all parties to pursue measures

[8] http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ECO/WKP(2018)63&docLanguage=En

domestically that will reduce greenhouse emissions. Countries are free to pursue measures of their choice but targets for emission reductions must be more ambitious than earlier ones. The members have agreed to track their progress and report this to each other, to co-operate and to provide support for climate action to developing countries. In June 2017, the United States announced its intention to withdraw from the Paris agreement with November 2020 the effective date of withdrawal.

TEST YOUR UNDERSTANDING 5.5

1 For each of the examples you provided in question 2(a) in Test your understanding 5.4, state and explain some method(s) that could be used to correct the externality.

2 Outline the three types of market-based policies that can be used to correct a negative production externality.

3 Using a diagram, show how a negative production externality created by the use of fossil fuels can be corrected by use of

- a taxes on output,
- b taxes on emissions,
- c tradable permits, and
- d legislation and regulation.
- e Discuss some advantages and disadvantages of each of these types of policy measures.

4 a Outline the meaning of and identify policies that 'internalise an externality'.

- b Compare and contrast market-based methods and command methods (such as legislation and regulations) to deal with negative production externalities.
- c Identify some difficulties that governments face in designing market-based methods.

5 a Define collective self-governance and discuss this as a method to deal with overuse of common pool resources.

- b Outline some possible limitations that this approach faces.

6 a Explain why taxes on emissions such as carbon taxes and tradable permits are similar with respect to their objectives. (*Hint:* think about incentives.)

- b Describe how they differ from taxes on output of the firms creating negative externalities.
- c Identify the policy that is preferable from the point of view of reducing the external costs of a negative production externality: a tax on output or a tax on emissions.

7 Using examples, explain under what circumstances international co-operation among governments is essential for the preservation of the environment.

REAL WORLD FOCUS 5.2

Effects of pesticides on bees

A pesticide known as sulfoxaflor is widely used to kill insects in agriculture. In 2013 it was approved by the Environmental Protection Agency (EPA) to be used in the United States. However, as sulfoxaflor was known to be toxic to bees, the EPA recommended that it should not be used when crops are flowering, which is when they attract bees.

Soon after, a number of beekeeping groups sued the EPA arguing that the use of sulfoxaflor would worsen the problem of the declining bee population. As a result, a US federal (national) court cancelled the EPA's approval of sulfoxaflor on the grounds that there was insufficient

Figure 5.8: Dead bees that died by pesticides

CONTINUED

evidence that it was not harmful to bees. In 2015 the EPA stopped allowing the sale of sulfoxaflor in the United States.

However, in 2016, the EPA reapproved the use of sulfoxaflor, but excluded crops that are attractive to bees and that flower at various times throughout the year. In addition, the EPA allowed the more general use of sulfoxaflor on an 'emergency' basis. The EPA has the authority to grant such exemptions even for pesticides that are not officially approved, in the event of a sudden appearance of insects that damage crops.

In the period 2017–2018, the EPA approved more than 100 'emergency exemptions' for sulfoxaflor. In 2018, it was used on over 16 million acres of farmland for cotton and sorghum alone.

Yet it is well known that the use of sulfoxaflor is harmful to bees, with serious consequences for their reproductive abilities. According to a senior scientist at the Center for Biological Diversity, 'Spraying 16 million acres of bee-attractive crops with a bee-killing pesticide in a time of global insect decline is beyond the pale* . . . The EPA is routinely misusing the "emergency" process to get sulfoxaflor approved because it's too toxic to make it through normal pesticide reviews.'

According to a study published in the journal *Biological Conservation*, there is a dangerous decline around the world in more than 40% of insect species and one third are endangered. The rate of extinction of insects is eight times faster than that of mammals, birds and reptiles. The cause is intensive agriculture and especially the heavy use of pesticides.

---

* This phrase means unacceptable, intolerable, outrageous

Source: *www.ecowatch.com; www.acs.org; www.theguardian.com*

Applying your skills

1 The extinction of insects represents a loss of ***biodiversity***, which is a common pool resource. Discuss how the concepts of rivalry and non-excludability can be used to explain depletion of this resource due to the heavy use of pesticides.

2 a Using a diagram explain what kind of externality is involved in the case of the harm being caused to bees.

b Identify and evaluate policies that could be used to deal with this externality.

c Using a diagram, show how this policy could work to correct the externality.

THEORY OF KNOWLEDGE 5.1

Economic thinking on sustainability and Elinor Ostrom, winner of the 2009 Nobel Prize in Economics

In our study of sustainability, we have seen that in economists' way of thinking, environmental destruction is analysed in terms of negative externalities. In the case of common pool resources, their overuse is simply the *external cost of* people's ordinary transactions that do not take account of consequences on the environment. In other words, the destruction of the environment is something that is thought of as being *outside the normal sphere of events*.

An important reason for thinking this way about the environment is that microeconomic theory is based on the idealised and fictional world of perfect markets with no failures, driven by economic decision-makers who make choices according to their best self-interest, responding to prices as signals and incentives, all of which leads to maximum social welfare.

CONTINUED

Economists' understanding of externalities as the divergence between private and social costs and private and social benefits is based on just this assumption of narrow self-interest: consumers and firms consume and produce taking only their own private interests into account, ignoring possible external costs that their actions give rise to. External costs arise *because consumers and producers are assumed to behave in ways that ignore the interests of society at large*.

The two main categories of measures to deal with environmental externalities we discussed, government regulations and market-based policies, are both based on this assumption about self-interested behaviour. Regulations use the command approach to force producers and consumers to reduce the external costs of their self-interested actions, and market-based policies create price incentives that try to bring the self-interested behaviour of economic decision-makers in line with society's best interests.

Yet all this raises some interesting questions. Does it make sense to view environmental destruction purely as the by-product of people's indifference toward the environment because they are self-interested beings? Are there situations when people do not display the narrow self-interested behaviour assumed by standard economic theory? If so, what are the implications, and what would be the appropriate policy responses?

These are the kinds of questions posed by Elinor Ostrom, who won the Nobel Prize in Economics (2009) for her work on the collective self-governance of common pool resources (discussed above). In one of her major works, Ostrom writes that her central question

> *'is how a group of principals who are in an interdependent situation can organise and govern themselves to obtain continuing joint benefits when all face temptations to free-ride, shirk, or otherwise act opportunistically.'*[9]

As discussed in the text, Ostrom's hundreds of experiments show that people can find solutions to managing resources sustainably by making rules regarding their behaviour and management of the common pool resources that must be followed by everyone. Her findings do not run entirely counter to the conclusions of standard economic theory, since lack of good methods of communication between people do result in the failures we studied in this chapter. However, her findings depart from standard economic theory in a very important respect: she has found that *these failures are not inevitable*, because rational economic behaviour does not always mean acting in one's own best self-interest; it often means acting in what is in the *group's best interests*. This kind of action arises as a result of institutions that permit communication between the resource users, leading to binding agreements, with monitoring and enforcement rules that ensure sustainable resource use.

Ostrom's conclusions focus on the point that people often behave *co-operatively rather than competitively,* and this has very important policy implications. Sometimes, the best method of preserving common pool resources is by allowing the resource users themselves to manage them, rather than through centralised government interference. When the conditions for co-operative solutions are present, the government's role should be to promote institutions that enable the users to manage the resource, such as a court system for resolving disputes and institutions that provide scientific knowledge for resource management. Often, this involves recognising that local co-operative institutions may be superior to institutions that are imposed on people from above or from outside.

If you decide to read further about Ostrom's work, you will find that she uses the term 'common pool resources'; other scholars use 'common property resources'. Both these terms mean common pool resources that are governed by a set of rules.

[9] Elinor Ostrom (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press.

CONTINUED

Thinking points

- How realistic do you think is the assumption that economic decision-makers are motivated by rational self-interest in making economic decisions?
- Many economists argue that even if rational self-interest is not a realistic assumption, it does not matter as long as the predictions of a theory fit with what happens in the real world (see also Theory of knowledge 7.1). What does Ostrom's work tell us about this perspective?
- If people sometimes behave co-operatively rather than competitively, what are the implications for the idea that environmental destruction is caused by externalities?
- Ostrom suggests that many people would change their behaviour if they understood that certain choices would be in their own best interests as well as in society's best interests (for example, biking rather than using a car). But many people may not know about such joint personal and social benefits. What can be done about this?

5.4 Negative consumption externalities

LEARNING OBJECTIVES

After studying this section, you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain negative externalities of consumption and the resulting welfare loss (AO2)
- draw a diagram illustrating negative externalities of consumption and welfare loss (AO4)
- explain the meaning of demerit goods (AO2)
- calculate welfare loss that arises from negative externalities of consumption (HL only) (AO4)
- explain government intervention to correct negative externalities of consumption: (AO2)
  - indirect (Pigouvian) taxes
  - legislation and regulation
  - education and awareness creation
  - nudges (HL only)
- draw diagrams to illustrate the above government responses (AO4)
- discuss strengths and limitations of the above government policies with respect to: (AO3)
  - difficulties in measurement of externalities
  - degree of effectiveness
  - consequences for stakeholders

Explaining and illustrating negative consumption externalities

Negative consumption externalities refer to external costs created by consumers. For example, when consumers smoke in public places, there are external costs that spill over onto society in the form of costs to non-smokers due to passive smoking. In addition, smoking-related diseases result in higher than necessary health care costs that are an additional burden upon society. When there is a consumption externality, the marginal private benefit (demand) curve does not reflect social benefits. In Figure 5.9, the buyers of cigarettes have a demand curve, *MPB*, but, when smoking, create external costs for non-smokers. These costs can be thought of as 'negative benefits', which therefore cause the *MSB* curve to lie below the *MPB* curve. The vertical difference between *MPB* and *MSB* represents the external costs. Note that since the externality involves consumption (i.e. the demand curve), the supply curve represents both marginal private costs and marginal social costs. The market determines an equilibrium quantity, Q_m, and price P_m, given by the intersection of the *MPB* and *MPC* curves, but the social optimum is Q_{opt} and P_{opt}, determined by the intersection of the *MSB* and *MSC* curves.

Other examples of negative consumption externalities include heating homes and driving cars by use of fossil fuels that pollute the atmosphere.

When there is a negative consumption externality, the free market overallocates resources to the production of the good, and too much of it is produced relative to what is socially optimum. This is shown by $Q_m > Q_{opt}$ and *MSC* > *MSB* at Q_m in Figure 5.9.

In general, negative externalities, whether these arise from production or consumption activities, lead to allocative inefficiency arising from an overallocation of resources to the good and to its overprovision.

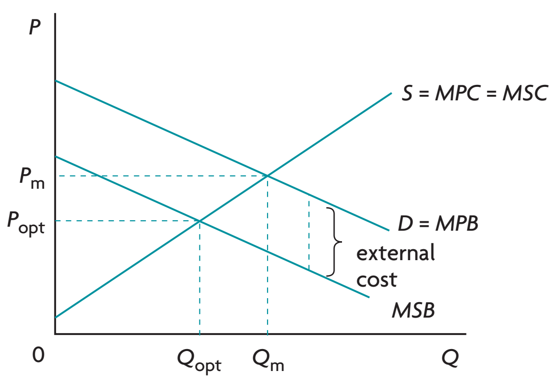


Figure 5.9: Negative consumption externality

The welfare loss of negative consumption externalities

Welfare loss

The welfare loss resulting from negative consumption externalities is the shaded area in Figure 5.10(a) and represents the reduction in benefits for society due to the overallocation of resources to the production of the good. For all units of output greater than Q_{opt}, *MSC* > *MSB*, indicating that too much of the good is produced. The welfare loss is equal to the difference between the *MSC* and *MSB* curves for the amount of output that is overproduced relative to the social optimum ($Q_m - Q_{opt}$). It represents the loss of social benefits from overproduction due to the externality. If this externality were corrected, society would gain the benefits represented by the shaded area. Note that, once again, the point of the welfare loss triangle lies at the Q_{opt} quantity of output (as in the case of negative production externalities).

Calculating welfare loss (HL only)

Figure 5.10(b) is the same as part (a) only with figures so we can calculate welfare loss. The area of the shaded triangle is the height times the width of the triangle divided by 2. Note that the height of the triangle is equal to the external cost per unit, or MPB – MSB, and the width is equal to the amount of overproduction by the market or $Q_m - Q_{opt}$:

$$\text{Welfare loss} = \frac{(6-4)\times(100-70)}{2} = \frac{2\times 30}{2} = \$30$$

Welfare loss in relation to consumer and producer surplus (Supplementary material)

Figure 5.10(c) shows how the welfare loss of a negative consumption externality is related to consumer and producer surplus and the external cost. In market equilibrium, consumer surplus is equal to the areas a + b, while producer surplus is equal to the areas c + d + f. The cost of the externality is represented by a + d + e (it is the difference between the *MPB* and *MSB* curves up to Q_m). The total social benefits are therefore consumer surplus plus producer surplus minus the external cost:

$$(a + b) + (c + d + f) - (a + d + e) = b + c + f - e$$

At the social optimum, consumer surplus is equal to b + c, and producer surplus is equal to f, while external costs are zero. Therefore, the total social surplus is equal to producer plus consumer surplus:

$$b + c + f$$

Comparing the total social benefits at market equilibrium and at the social optimum, we see they are smaller at market equilibrium by the area e, which is the welfare loss.

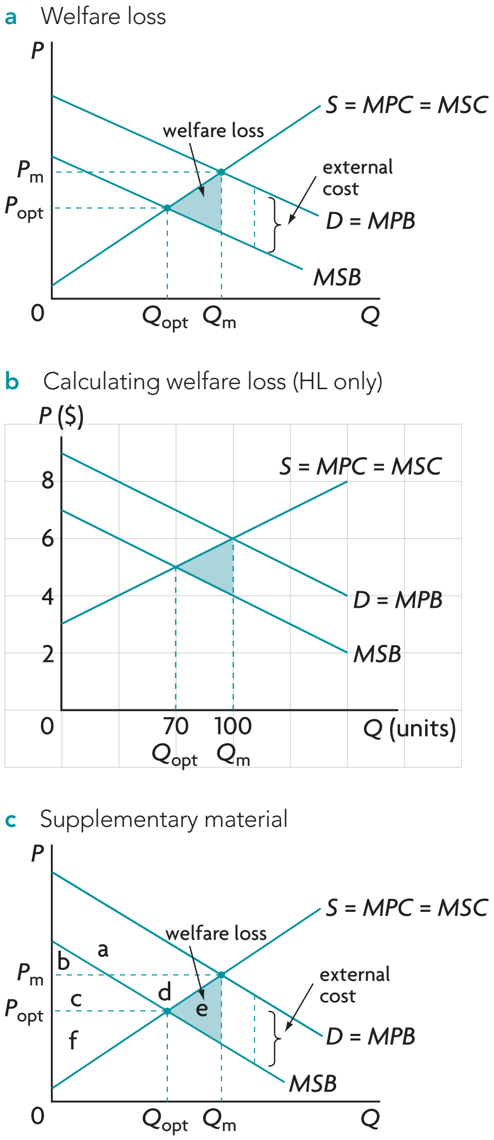


Figure 5.10: Welfare loss in a negative consumption externality

The case of demerit goods

Demerit goods are goods that are considered to be undesirable for consumers, but which are overprovided by the market. Examples of demerit goods include cigarettes, alcohol and gambling. One important reason for overprovision is that the good may have negative consumption externalities, in which case the market overallocates resources to its production. This could occur because of consumer ignorance about its negative effects or indifference: consumers may not be aware of the harmful effects upon others of their actions, or they may not care.

Negative consumption externalities and overuse of common pool resources

Overuse of common pool resources is seen as resulting more from production activities than consumption activities, therefore relating more to negative production rather than consumption externalities. However, overuse of common pool resources also results from negative consumption externalities, shown in Figure 5.9. Take the demand for heating oil, represented by the demand curve *MPB*. The overuse of clean air (the common pool resource) is the external cost that causes the marginal social benefit curve (*MSB*) to lie below the *MPB* curve. The same applies to the use of cars that run on gasoline (petrol). More importantly, air travel is a consumption activity that results in significant and rapidly increasing greenhouse gases that cause global warming. These activities on the part of consumers result in overuse of the common pool resource of clean air, seriously threatening the stability of the global climate.

TEST YOUR UNDERSTANDING 5.6

1. a. Using a diagram, show how marginal private benefits and marginal social benefits differ when there is a negative consumption externality.
   b. Explain the difference between the equilibrium quantity determined by the market and the quantity that is optimal from the point of view of society's preferences.
   c. Describe the problem with the allocation of resources achieved by the market when there is a negative consumption externality.
   d. Show the welfare loss created by the negative consumption externality in your diagram, and explain what this means.
2. Provide some examples of negative consumption externalities.

Policies to correct negative consumption externalities and prevent overuse of common pool resources

Market-based policies

A key market-based policy to correct negative consumption externalities (including demerit goods) involves the imposition of indirect Pigouvian taxes, as in the case of negative production externalities. Indirect taxes can be imposed on the good whose consumption creates external costs (for example, cigarettes and petrol/gasoline).

The effects of an indirect tax are shown in Figure 5.11(a). When such a tax is imposed on the good whose consumption creates the external cost, the result is a decrease in supply and an upward shift of the supply curve from *MPC* to *MPC* + tax. If the tax equals the external cost, the *MPC* + tax curve intersects *MPB* at the Q_{opt} level of output, and quantity produced and consumed drops to Q_{opt}. (The demand curve does not shift but remains at $D = MPB$.) Q_{opt} is the socially optimum quantity, and price increases from P_m to P_c. The tax therefore permits allocative efficiency to be achieved.[10]

Advantages and disadvantages

As with negative production externalities, economists prefer market-based solutions to the problem of negative consumption externalities, as long as the situation permits the use of market-based policy. Therefore, indirect taxes are the preferred measure, as they internalise the externality (as in the case of negative production externalities). By changing relative prices, indirect taxes create incentives for consumers to change their consumption patterns; the good that is taxed becomes relatively more expensive and consumption is reduced.

However, there are a number of difficulties in this approach. The first involves difficulties in measuring the value of the external costs. Take, for example, the case of passive smoking, an external cost created by smokers, or the case of petrol (gasoline) consumption, which creates external costs in the form of environmental pollution. There are many technical difficulties involved in trying to assess who and what is affected, as well as to determine the value of the external costs, on the basis of which a tax can be designed.

A further difficulty is that some of the goods whose consumption leads to negative consumption externalities (for example, petrol/gasoline and cigarettes) have an inelastic demand. As you may remember from Chapter 3, when demand is inelastic, the percentage decrease in quantity demanded is smaller than the percentage increase in price (due to the tax). Therefore, it is possible that imposing taxes on such goods as petrol (gasoline) and cigarettes (both of which have an inelastic demand) works to increase government tax revenues while not significantly decreasing the quantity demanded of these goods. This could mean that in order to achieve Q_{opt}, a very high indirect tax would have to be imposed, which would very likely be politically unacceptable. On the other hand, the large tax revenues can be used to finance education programmes to discourage consumption of particular goods.

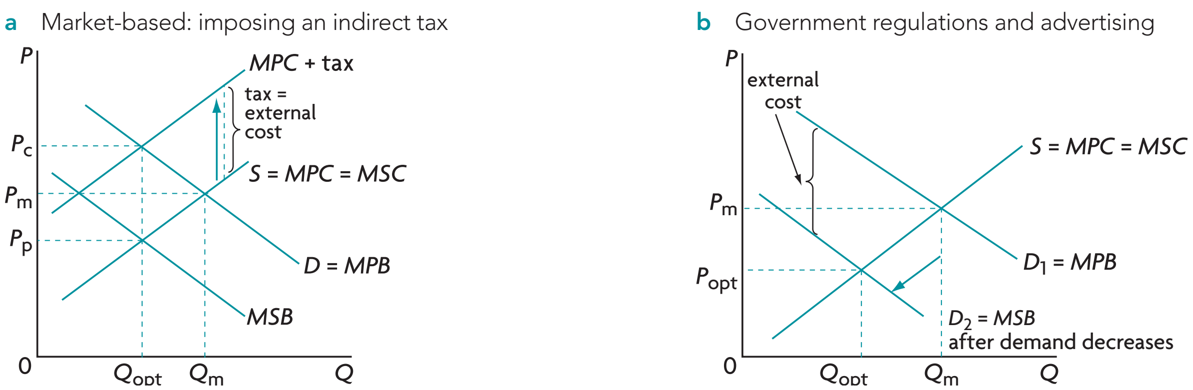


Figure 5.11: Correcting negative consumption externalities

[10] Note that the new equilibrium price, P_c, is the price paid by consumers; the price received by producers is $P_p = P_c$ minus tax per unit (see Chapter 4, section 4.2).

A tax on producers or consumers? (Supplementary material)

You may be wondering why the tax to correct negative consumption externalities should affect producers, shifting the supply curve upward in Figure 5.11(a), and not consumers by shifting the demand curve downward, who after all are the ones creating the externality through consumption. You will find the answer to this question in the digital coursebook as Supplementary material.

Government legislation and regulation

If negative consumption externalities were corrected, Q_{opt} quantity of the good would be produced, reflecting allocative efficiency. Regulations can be used to prevent or limit consumer activities that impose costs on third parties, such as legal restrictions on activities like smoking in public places or age restrictions forcing sellers to do business only with adults. This has the effect of shifting the $D_1 = MPB$ curve towards the MSB curve in Figure 5.11(b), until D_2 overlaps with MSB. This would eliminate the externality, with production and consumption occurring at Q_{opt} and price falling to P_{opt}.

Advantages and disadvantages

Some regulations can be very effective in reducing the external costs of certain consumption activities. Examples include prohibiting smoking in public places and placing restrictions on cars entering city centres. However, they cannot be used to deal with other kinds of negative consumption externalities. For example, it would be very difficult to regulate petrol (gasoline) consumption; on the other hand, imposing indirect taxes on such goods may be more effective (though subject to the limitations noted above on inelastic demand).

Education and awareness-creation

Educating the public and creating awareness by the government can be used to try to persuade consumers to buy fewer goods with negative externalities, such as anti-smoking campaigns, campaigns to avoid consumption of unhealthy foods, or campaigns to reduce the consumption of goods and services based on fossil fuel use. Examples include using public transportation to economise on petrol (gasoline) use or improving home insulation to reduce oil consumption (for heating) or providing information to consumers on the amount of carbon produced by air travel.

The objective is to try to decrease demand for goods giving rise to negative consumption externalities, and the effects are the same as with government regulations, shown in Figure 5.11(b). The MPB curve shifts to D_2 after the campaign, where it coincides with MSB, where Q_{opt} is produced and consumed, and the price falls from P_m to P_{opt}.

Advantages and disadvantages

Education and creation awareness have the advantage that they are simpler than other methods, but they too have their disadvantages. One of these involves the cost to the government of campaigns, which are funded out of tax funds, meaning there are less funds available for use elsewhere in the economy (there are opportunity costs). There is also the possibility that such methods may not be effective enough in reducing the negative externality.

Nudges (HL only)

Nudges were introduced in Chapter 2 where we saw that they form a part of behavioural economics, involving the design of methods to influence consumers' behaviour. They can be used in ways similar to education and consumer awareness to encourage consumers to rely less on goods with negative externalities, such as unhealthy foods and cigarettes. For example, unhealthy foods can be placed in less accessible places in shops, or graphic pictures of the consequences of smoking may be placed on cigarette packs. In such cases demand for the product falls so that the MPB curve shifts toward MSB as in Figure 5.11(b). In addition, nudges can be used to encourage desirable behaviour, such as creating bicycle lanes to motivate car drivers to use bicycles instead.

Advantages and disadvantages

There may be difficulties in designing effective nudges as these are still a new method intended to influence consumer behaviour, and not enough is known yet about how consumers respond to particular nudges and choice architecture. There is, too, the issue that particular nudges may not have the same effects across income groups or cultural groups.

> Correction of negative consumption externalities involves either decreasing supply and shifting the *MPC* curve upward by imposing an indirect (Pigouvian) tax; or by decreasing demand and shifting the *MPB* curve toward the *MSB* curve through regulations, education and awareness creation or nudges. Both supply decreases and demand decreases are intended to lead to production and consumption at Q_{opt} and the achievement of allocative efficiency.

In general, given the limitations above for all the various policies it is only possible to move the economy in a direction towards correction of the externality, rather than achieving a precise allocation of resources where Q_{opt} is produced and consumed. Very often, the use of several policies together may be the best way to address the externality. For example, tax revenues from the imposition of a Pigouvian tax may be used to finance education and awareness creation programmes as well as nudges. Governments must therefore be selective in the methods they use to reduce consumption externalities, depending on the particular good that creates the external costs.

TEST YOUR UNDERSTANDING 5.7

1. For each of the examples you provided in question 2 of Test your understanding 5.6, explain some methods that could be used to correct the externality.
2. Using diagrams, show how a negative consumption externality can be corrected by use of
   a. legislation and regulations that limit the external costs,
   b. education and awareness creation,
   c. (HL only) nudges, and
   d. indirect taxes.
   e. Identify some advantages and disadvantages of each of these policy measures.
3. Explain the meaning of a demerit good and provide examples.
4. Explain how a negative consumption externality differs from a negative production externality.
5. a. Explain what kinds of measures economists prefer to correct negative consumption externalities.
   b. Identify situations where these might not be very effective.

REAL WORLD FOCUS 5.3

Policies to reduce sugar consumption

Following a 2015 report by the World Health Organization (of the United Nations) detailing the negative health effects of sugar, more and more countries around the world are imposing a tax on the sugar content of foods and drinks in an effort to reduce obesity and other health problems like diabetes arising from sugar overconsumption. These taxes have generally met with little opposition from the public. The taxes are imposed on foods and/or drinks with sugar, which are taxed in accordance with the quantity of sugar they contain.

Figure 5.12: Classic coke food label decoder showing high sugar level

In response to the implementation of these policies, many major food producers around the world have responded by creating new products with no added sugar, or reduced sugar, while at the same time trying to maintain the original taste. In some reduced sugar items, the sugar content is lowered enough so that the tax can be avoided altogether. In addition, some governments are working with food producers to establish voluntary agreements with producers to lower the sugar content of their food products.

Another approach involves banning the advertisement of unhealthy foods during the times when children watch television. A still different approach involves the use of electronic ordering systems that offer consumers healthy goods as the default options.

CONTINUED

Applying your skills

1. Using a diagram, explain what kind of externality the article is referring to.
2. Using a diagram, explain how a tax on the sugar content of foods and drinks can help reduce the externality.
3. Outline a likely reason that the taxes have generally met with little public opposition.
4. a. Identify and explain the other types of policies (other than indirect taxes) that the article refers to, that could help to reduce the externality.

   b. Using diagrams, show how these policies are expected to work.

Source: *www.kerry.com*

THEORY OF KNOWLEDGE 5.2

The ethical dimensions of sustainability and preserving the global climate

In Chapter 1 we saw that solutions to the problem of sustainability face major technical difficulties due to uncertainties and incomplete knowledge of social and natural scientists regarding the complex relationships between environmental, economic, social and institutional variables. These kinds of technical difficulties are also responsible for the uncertainties surrounding both regulatory and market-based economic policies to address environmental externalities discussed in the present chapter.

Over and above the technical difficulties, the problem of sustainability faces major ethical issues of fairness and justice, relating to intergenerational equity (running from generation to generation), as well as equity across nations and social groups within nations of the present generation.

In the area of climate change alone, important issues include (a) how will the burden of having to make sacrifices in the present be distributed among countries; (b) how will the impacts of climate change be evaluated; and (c) how will intergenerational equity be accounted for?[11]

To determine the distribution of sacrifices, a possible ethical principle that can be used is 'the polluter pays' principle, according to which the sacrifice is distributed according to how much each country contributes to climate change. In one variant of this principle, it would be necessary to take into account cumulative (historical) contributions to greenhouse gas emissions. This would place an extra burden on the developed countries of today, which over time, have contributed far more to emissions than developing countries. As a counterargument, opponents refer to 'excusable ignorance', meaning it should not be necessary to pay for past emissions if these were made without knowledge of their effects on the global climate. According to a different ethical principle, the past would be ignored and future emissions rights would be distributed to all countries on a *per capita* basis.

On the second issue, concerning evaluation of impacts of climate change, one approach involves welfare analysis. This has given rise to disagreements about how to calculate welfare and add it up across individuals in the present as well as in the future. Another approach focuses on human rights as the basis for evaluating impacts, such as the rights to food, water and shelter, which may be threatened by climate change.

[11] The World Bank (2009) *World Development Report 2010: Development and Climate Change.*

CONTINUED

Intergenerational equity, the third issue, is closely related to the evaluation of impacts of climate change, as these must account for impacts not only on the present generation but future generations as well.

These kinds of questions clearly belong to the normative realm of thought. Given the technical difficulties as well, it is no wonder that there are broad disagreements over sustainability, and no easy solutions appear on the horizon.

Thinking points

- What do you think should be the role of science and social science in providing answers to these kinds of questions?
- To what extent do you think market forces can be relied upon, if at all, to deal with problems of environmental sustainability?
- Market economies are based upon human behaviour motivated by rational self-interest (see Chapter 2). To what extent do you think this self-interest is the root cause of the environmental problems that beset the human race today? (See also Theory of knowledge 5.1.)
- Given that, historically, economically more developed countries have been mainly responsible for today's environmental problems, do you agree with the view that economically less developed countries should simply ignore calls for them to limit their growth rates to prevent further global warming?

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1. Identify a good or service in your country or a country you are interested in that gives rise to a particular negative production externality. Examine the extent of the external costs and how they affect various stakeholders. Investigate the solutions that are being used to reduce its production and consumption. Discuss the strengths and limitations of these solutions, and try to identify the effects on various stakeholders.
2. See question 1 above and apply it to a negative consumption externality.
3. Identify a scheme of tradable permits that has been implemented. Research and discuss its strengths and limitations.
4. Identify a common pool resource that is of interest to you and investigate what, if any, measures are being taken in your country of residence or a country you are interested in to deal with its possible overuse.
5. As indicated in Real world focus 5.3, a number of countries around the world have imposed or are planning to impose indirect (Pigouvian) taxes on various food items (sugar and/or fats) that are considered to be harmful to human health. Examine whether such a tax is used or is under consideration in your country or a country you are interested in, and examine the external costs of consumption of such foods, as well as the effects of the tax on various stakeholders. Investigate whether any studies have been done to evaluate the impact of the tax, and the extent to which it has been successful with respect to attaining its objective of reducing consumption of such foods. Consider also additional, complementary policies that are in use for this purpose in your country of choice.

CONTINUED

6 Air travel is the fastest-growing contributor to carbon emissions. Responsible for about 5% of global warming in 2018, by 2050 emissions may grow by more than 700%, taking up one-quarter of the global carbon budget for limiting the temperature rise to 1.5%. Yet the global airline industry rejects policies such as carbon pricing and regulation of aviation emissions by governments or international bodies. Instead it has opted for a scheme called Carbon Offsetting and Reduction Scheme for International Aviation (CORSIA) which has been criticised as being ineffective. (a) Research CORSIA and identify some advantages and disadvantages as a method to tackle the problem of airline emissions. (b) Research and outline steps that travellers can take in order to reduce their carbon footprint when they travel by air.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 6

Market failure and socially undesirable outcomes II:

Positive externalities, public goods, asymmetric information and inability to achieve equity

BEFORE YOU START

- Can you think of some goods or services that when consumed or produced bring benefits to others even though they didn't consume or produce them?
- You may have noticed that some goods and services are provided free of charge, such as roads, parks, street lighting, public schools, and many more. Who pays for these and what might be the reasons people can enjoy them without having to pay for them?

In this chapter we continue the discussion begun in Chapter 5 on the inability of the market to fulfill some of its promises. We will begin by examining positive externalities of production and consumption. We will then discover public goods, which are not produced at all by the market even though they are socially desirable. We will then examine the consequences of information asymmetries between buyers and sellers, and will conclude with a discussion of the market's inability to achieve equity in the distribution of income and wealth.

Photo: Tooth wheel mechanism

6.1 Positive production externalities

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain positive externalities of production and the resulting welfare loss (AO2)
- draw a diagram illustrating positive externalities of production and welfare loss (AO4)
- calculate welfare loss that arises from positive externalities of production (HL only) (AO4)
- explain government intervention to correct positive externalities of production: (AO2)
    - government provision
    - subsidies
- draw diagrams to illustrate the above government responses (AO4)
- discuss strengths and limitations of the above government policies with respect to: (AO3)
    - difficulties in measurement of externalities
    - degree of effectiveness
    - consequences for stakeholders

Explaining and illustrating positive production externalities

Externalities as a form of market failure were introduced in Chapter 5, where we discussed negative externalities of production and consumption. We now turn our attention to positive externalities.

Positive production externalities refer to external benefits created by producers. If, for example, a firm engages in research and development, and succeeds in developing a new technology that spreads throughout the economy, there are external benefits because not only the firm but also society benefits from widespread adoption of the new technology. Therefore, the social costs of research and development are lower than the private costs. In Figure 6.1, the *MSC* curve lies below the *MPC* curve, and the difference between the two curves is the value of the external benefits (these can be thought of as 'negative costs'). The demand curve represents both *MPB* and *MSB* since the externality involves only production. The market gives rise to equilibrium quantity Q_m and price P_m, determined by the intersection of the *MPB* and *MPC* curves, while the social optimum is given by Q_{opt} and P_{opt}, determined by the intersection of the *MSB* with *MSC* curves. Since $Q_m < Q_{opt}$, the market underallocates resources to research and development activities that lead to new technologies, and not enough of them are undertaken.

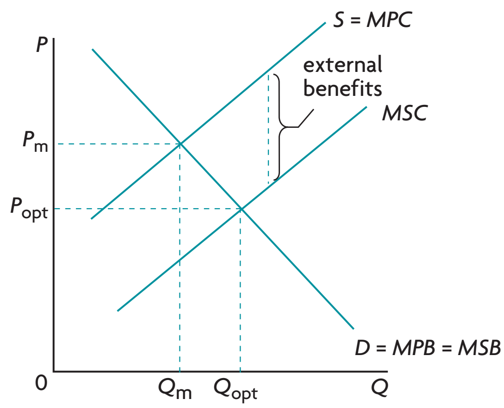


Figure 6.1: Positive production externality

When there is a positive production externality, the free market underallocates resources to the production of the good: too few resources are allocated to its production, and too little of it is produced. This is shown by $Q_m < Q_{opt}$ and $MSB > MSC$ at Q_m in Figure 6.1.

More examples of positive production externalities include:

- firms train workers who later switch jobs and work elsewhere; external benefits are created as the new employers and society benefit from the trained workers
- a firm provides first aid classes to employees to improve work safety; external benefits are created as this knowledge is applied also to people outside the workplace.

The welfare loss of positive production externalities

Welfare loss

The underallocation of resources to the production of a good with a positive production externality leads to a welfare loss, shown in Figure 6.2(a) as the shaded area. This loss is equal to the difference between the *MSB* and *MSC* curves for the amount of output that is underproduced relative to the social optimum ($Q_{opt} - Q_m$). It involves external benefits for society that are lost because not enough of the good is produced. If the externality were corrected, society would gain the benefits represented by the shaded area. Note that the point of the welfare loss triangle lies at the Q_{opt} quantity of output.

Calculating welfare loss (HL only)

Figure 6.2(b) is similar to part (a) with figures so we can calculate welfare loss. The area of the welfare loss triangle is the height times the width of the triangle divided by 2. The height of the triangle is equal to the external benefit per unit, or *MPC* – *MSC*, and the width is equal to the amount of underproduction by the market or $Q_{opt} - Q_m$:

$$\text{Welfare loss} = \frac{(6-4)\times(90-60)}{30} = \frac{2\times 30}{2} = \$30$$

Welfare loss in relation to consumer and producer surplus (Supplementary material)

Figure 6.2(c) shows the welfare loss in relation to consumer and producer surplus and the externality. At market equilibrium, consumer surplus is area a, producer surplus is area b + e, and the external benefits are c + f (the difference between the *MPC* and *MSC* curves up to the point of production by the market, Q_m). The total benefits are therefore consumer surplus plus producer surplus plus external benefits:

a + (b + e) + (c + f) = a + b + e + c + f

At the social optimum, consumer surplus is a + b + c + d, producer surplus is e + f + g, and external benefits are zero, making a total of:

(a + b + c + d) + (e + f + g) = a + b + c + d + e + f + g

Comparing total benefits at market equilibrium and at the social optimum, we find that at the social optimum there are additional benefits of the amount d + g, corresponding to the shaded area in the figure. This is the amount of welfare that is lost at market equilibrium due to underallocation of resources arising from the positive production externality.

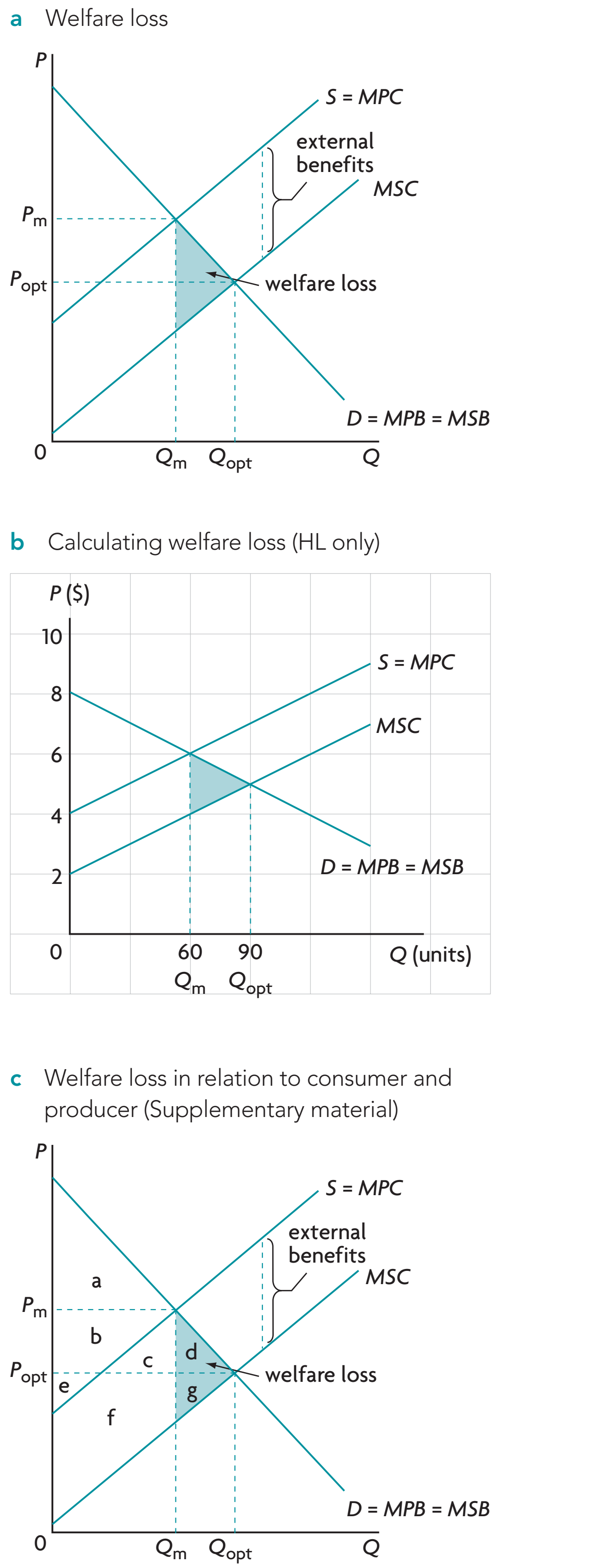


Figure 6.2: Welfare loss in a positive production externality

Correcting positive production externalities

Direct government provision

A solution often pursued by governments involves direct government provision of the good or service creating the positive production externality. For example, governments often engage in research and development (R&D) for new technology, for medicine and pharmaceuticals, and many other areas. The government can also directly provide training for workers. Governments pay for such activities with government funds, raised through taxes. Figure 6.3(a) shows that when the government intervenes by providing goods and services itself, this has the effect of shifting the supply curve (= *MPC* curve) downward (or to the right), toward the *MSC* curve so that the optimum quantity of the good, Q_{opt}, will be produced, with price falling from P_m to P_{opt}.

Subsidies

We studied subsidies and their effects in Chapter 4, where we saw how their introduction into a perfect market (with no market failures) creates allocative inefficiency. Now, we will see how subsidies can correct allocative inefficiency by correcting a market failure.

If the government provides a subsidy to a firm per unit of the good produced that is equal to the external benefit, then the marginal private cost (*MPC* = supply) curve shifts downward (or rightward[1]) until it coincides with the *MSC* curve, as shown in Figure 6.3(b). The result is to increase quantity produced to Q_{opt} and to lower the price from P_m to P_{opt}. The problem of underallocation of resources and underprovision of the good is corrected, and allocative efficiency is achieved.

You may note that direct government provision and subsidies have the same market outcomes.

> Correction of positive production externalities involves shifting the *MPC* curve downward toward the *MSC* curve through direct government provision or by subsidies. For allocative efficiency to be achieved, the quantity produced and consumed must increase to Q_{opt} as price falls to P_{opt}.

Evaluating policies to correct positive production externalities

This topic will be discussed together with policies to correct positive consumption externalities below (Section 6.2) because of similarities of the policies involved.

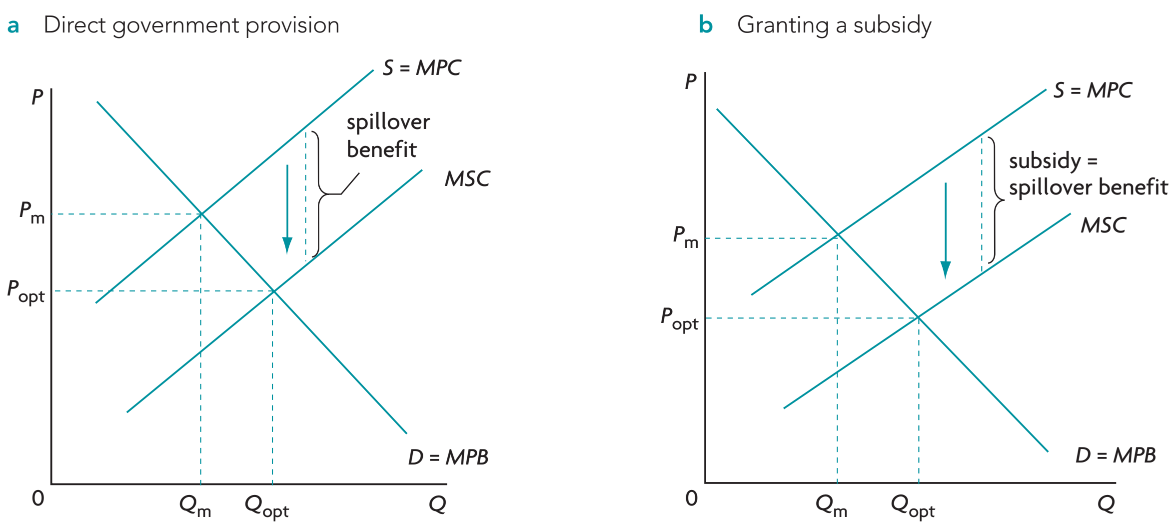


Figure 6.3: Correcting positive production externalities

[1] See 'Quantitative techniques' in the digital coursebook for an explanation of the equivalence of downward and rightward shifts of the supply curve.

TEST YOUR UNDERSTANDING 6.1

1 a Using a diagram, show how marginal private costs and marginal social costs differ when there is a positive production externality.

b Explain the difference between the equilibrium quantity determined by the market and the quantity that is optimal from the point of view of society's preferences.

c Describe what this tells you about the allocation of resources achieved by the market when there is a positive production externality.

d Show the welfare loss created by the positive production externality in your diagram, and explain what this means.

2 Provide some examples of positive production externalities.

3 For each of the examples you provided in question 2, explain some methods that can be used to correct the externality.

6.2 Positive consumption externalities

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain positive externalities of consumption and the resulting welfare loss (AO2)
- draw a diagram illustrating positive externalities of consumption and welfare loss (AO4)
- explain the meaning of merit goods (AO2)
- calculate welfare loss that arises from positive externalities of consumption (HL only) (AO4)
- explain government intervention to correct positive externalities of consumption: (AO2)
  - legislation and regulation
  - education and awareness creation
  - nudges (HL only)
  - government provision
  - subsidies
- draw diagrams to illustrate the above government responses (AO4)
- discuss strengths and limitations of the above government policies with respect to: (AO3)
  - difficulties in measurement of externalities
  - degree of effectiveness
  - consequences for stakeholders

Explaining and illustrating positive consumption externalities

When there is a **positive consumption externality**, external benefits are created by consumers. For example, the consumption of education benefits the person who receives the education, but in addition gives rise to external benefits, involving social benefits from a more productive workforce, lower unemployment, higher rate of growth, more economic development, lower crime rate, and so on. Similarly, the consumption of health care services benefits not only the person receiving the services but also society and the economy, because a healthier population is more productive, enjoys a higher standard of living, does not pass on contagious diseases as much and may have a higher rate of economic growth. In Figure 6.4, we see that the marginal social benefit (*MSB*) curve lies above the marginal private benefit (*MPB*) curve, and the difference between the two consists of the external benefits to society. The socially optimum quantity, Q_{opt}, is given by the point where $MSB = MSC$, and the quantity produced by the market is given by the point where $MPB = MPC$. Since $Q_{opt} > Q_m$, the market underallocates resources to the good or service, and too little of it is produced.

When there is a positive consumption externality, the free market underallocates resources to the production of the good, and too little of it is produced relative to the social optimum. This is shown by $Q_m < Q_{opt}$ and $MSB > MSC$ at Q_m in Figure 6.4.

In general, positive externalities (external benefits), whether these arise from production or consumption activities, lead to an underallocation of resources to the good in question, and therefore to its underprovision.

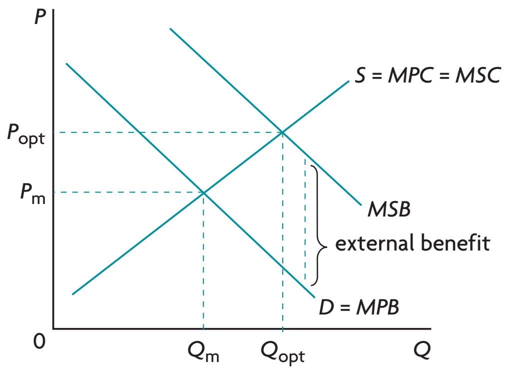


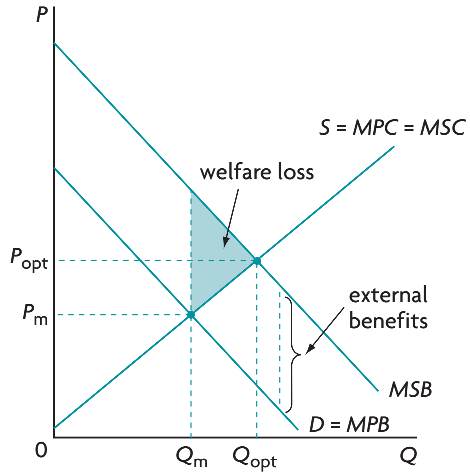
Figure 6.4: Positive consumption externality

The welfare loss of positive consumption externalities

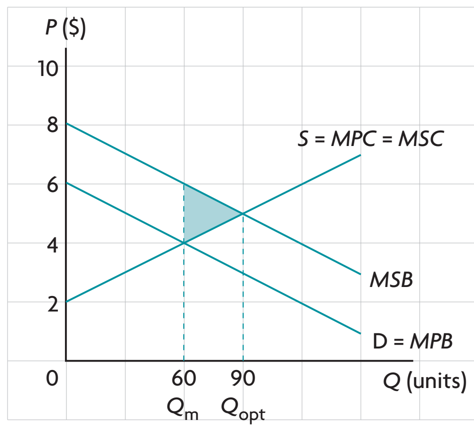
Welfare loss

The welfare loss arising from a positive consumption externality is the shaded area in Figure 6.5(a), and is the difference between the *MSB* and *MSC* curves for the amount of output that is underproduced relative to the social optimum ($Q_{opt} - Q_m$). It represents the loss of social benefits due to underproduction of the good. If this externality were corrected, society would gain the benefits represented by the shaded area. Once again, we see that the point of the welfare loss triangle lies at the Q_{opt} quantity of output.

a Welfare loss



b Calculating welfare loss (HL only)



c Welfare loss in relation to consumer and producer surplus (Supplementary material)

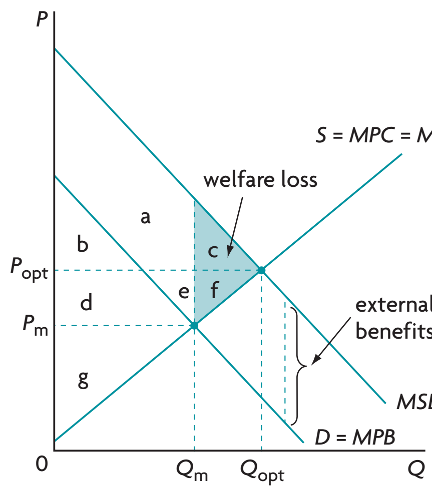


Figure 6.5: Welfare loss in a positive consumption externality

You may have noticed that in the case of negative externalities of production and consumption, where $Q_{opt} < Q_m$, the welfare loss triangle always points leftward toward Q_{opt}. By contrast, in the case of positive externalities of production and consumption, where $Q_{opt} > Q_m$, the welfare loss triangle always points rightward, again toward Q_{opt}.

Calculating welfare loss (HL only)

Figure 6.5(b) is similar to part (a) with figures so we can calculate welfare loss. The area of the welfare loss triangle the height times the width of the triangle divided by 2. The height of the triangle is equal to the external benefits per unit, or $MSB - MPB$, and the width is equal to the amount of underproduction by the market or $Q_{opt} - Q_m$:

$$\text{Welfare loss} = \frac{(6-4)\times(90-60)}{30} = \frac{2\times 30}{2} = \$30$$

Welfare loss in relation to consumer and producer surplus (Supplementary material)

In Figure 6.5(c) we see how the welfare loss arises in relation to consumer and producer surplus and the external benefits. In market equilibrium, consumer surplus is equal to areas b + d, producer surplus is area g, and the external benefits are a + e (or the difference between MSB and MPB up to production at Q_m by the market). The total social benefits in market equilibrium are equal to consumer surplus plus producer surplus plus the external benefits:

$$(b + d) + g + (a + e) = b + d + g + a + e$$
$$= a + b + d + e + g$$

At the social optimum, consumer surplus is given by a + b + c, producer surplus is d + e + f + g, and the external benefits are zero. Therefore the total social benefits are:

$$(a + b + c) + (d + e + f + g) = a + b + c + d + e + f + g$$

Comparing the total social benefits at market equilibrium with those at the social optimum, we find that at the social optimum they are greater by the amount c + f. This is the welfare loss that arises when production occurs at market equilibrium as a result of an underallocation of resources due to the positive consumption externality.

The case of merit goods

Merit goods are goods that are held to be desirable for consumers, but which are underprovided by the market. (Note that the term 'good' in the expression 'merit good' applies to both goods and services.) Reasons for underprovision include:

- **The good may have positive externalities.** In this case too little is provided by the market. Examples of merit goods include education (for the reasons noted above in the discussion of externalities); immunisation programmes (which benefit not only those who have received them but also the broader population by wiping out a disease).
- **Low levels of income and poverty.** Some consumers may want certain goods or services but cannot afford to buy them. Remember demand shows the quantities of a good or service that consumers are willing and able to buy at different prices. If they have low incomes, they may be willing but not able to buy something, in which case their desire does not show up in the market, and market demand (the sum of all individual demands) is too low. Examples include health care services, medicines, education and recreational facilities, which people on low incomes often cannot afford to buy in the market.
- **Consumer ignorance.** Consumers may be better off if they consume certain goods and services but they may be ignorant of the benefits, and so do not demand them. For example, preventive health care (such as immunisation, annual health check-ups) can prevent serious diseases, but lack of knowledge about the benefits may lead consumers to demand too little of these services.

Note that more than one factor may be at work simultaneously; for example, the underprovision of health care services can result from all three reasons listed above.

Correcting positive consumption externalities

Government legislation and regulation

Legislation can be used to promote greater consumption of goods with positive externalities. For example, most countries have legislation that makes education compulsory up to a certain age (note that education is a merit good). In this case, demand for education increases, and the demand curve $D_1 = MPB$ shifts to the right (or upward), as in Figure 6.6(a). Ideally, it will shift until it reaches the MSB curve, where $D_2 = MSB$, and Q_{opt} is produced and consumed.

Education and awareness creation

Governments can use education of the public, awareness creation, to try to persuade consumers to buy more goods with positive externalities. For example, they can

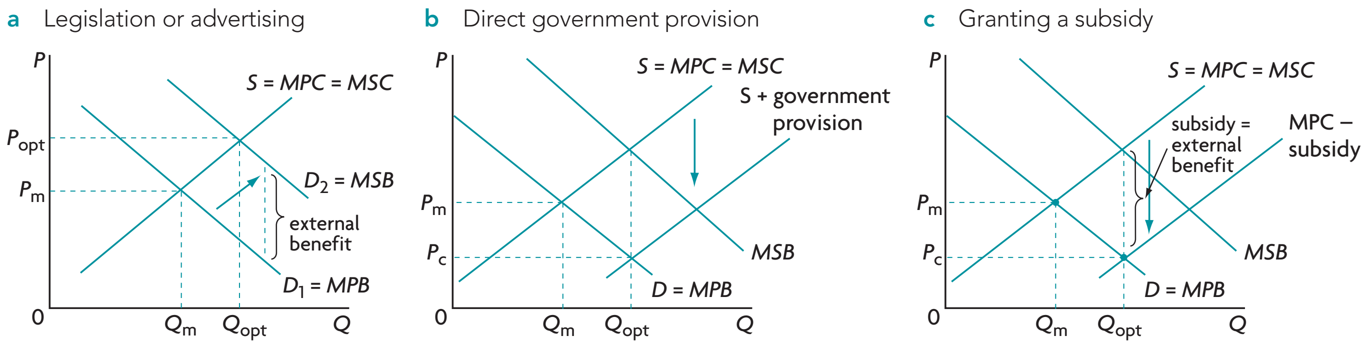


Figure 6.6: Correcting positive consumption externalities

try to encourage the use of sports facilities for improved health. The objective is to increase demand for such services, and the effect is the same as with legislation, shown in Figure 6.6(a): D_1 shifts to $D_2 = MSB$ and Q_{opt} is produced and consumed, while price increases to P_{opt}.

Nudges (HL only)

Nudges have similar effects as education and awareness creation programmes. Governments can use nudges like the creation of bicycle lanes to encourage the use of bike riding for more physical exercise. (Note that this nudge also has the effect of reducing the negative externality of gasoline (petrol) consumption from the use of cars, noted earlier.) Here, too, the objective is to increase demand shifting D_1 toward D_2.

Direct government provision

Governments are frequently involved in the direct provision of goods and services with positive consumption externalities. The most important examples include government (public) provision of education and health care in virtually all countries in the world. Education and health care are merit goods with external benefits so large and important that it is widely believed that they must not be left to private sector provision alone. In most countries where there is government provision of health care and education, there is also private sector provision of these services (though to varying degrees).

Direct government provision is shown in Figure 6.6(b), and has the effect of increasing supply and therefore shifting the supply curve *S* rightward (or downward) to *S* + government provision. To achieve the social optimum Q_{opt}, the new supply curve must intersect *MPB* at the level of output Q_{opt}, as seen in the figure. At the new equilibrium, price falls to P_c, Q_{opt} is produced and allocative efficiency is achieved.

Subsidies

A subsidy to the producer of the good with the positive externality has the same effects as direct government provision. It results in increasing supply and shifting the supply curve rightward (or downward), as shown in Figure 6.6(c) (which is similar to Figure 6.6(b)). If the subsidy is equal to the external benefit, the new supply curve is *MPC* – subsidy,[2] and it intersects *MPB* at the Q_{opt} level of output. Again, price falls from P_m to P_c, Q_{opt} is produced and allocative efficiency is achieved.

Correction of positive consumption externalities involves either increasing demand and shifting the *MPB* curve towards the *MSB* curve through legislation or education and awareness creation; or increasing supply and shifting the *MPC* curve downward by direct government provision or by granting a subsidy. Both demand increases and supply increases can lead to production and consumption at Q_{opt} and the achievement of allocative efficiency. The price paid by consumers increases when demand increases, and falls when supply increases.

Note that the problem of underprovision of merit goods by the market (defined above) can be addressed by all the methods noted above: legislation, education and awareness

[2] You may note that the reason for the minus sign in *MPC* – subsidy is that the amount of the subsidy is *subtracted* from *MPC* in order to arrive at the new supply curve.

creation, direct government provision and granting of subsidies. All are intended to increase the amount of the good produced and consumed, as increased consumption of such goods is held to be desirable for society.

Evaluating policies to correct positive production and consumption externalities

Both direct government provision and subsidies are widely used as methods to deal with positive consumption externalities, and to a lesser extent also with positive production externalities. Both methods are very effective in increasing the quantity of the good produced and consumed, and both have the added advantage of lowering the price of the good to consumers.

There are, however, difficulties involved in achieving the optimum results (where $MSC = MSB$). First, both direct government provision and granting of subsidies involve the use of government funds that rely on tax revenues. Governments generally have very many possible alternative uses for these funds, each of which has an opportunity cost. As it is not possible for the government to directly provide or subsidise all goods and services with positive externalities, choices must be made on (a) which goods should be supported, and (b) by how much they should be supported.

Ideally, choices should be made on the basis of economic criteria, which would specify the amount of social benefits expected in relation to the cost of providing them, the objective being to maximise the benefits for each good and service to be provided or subsidised for a given cost. However, in practice it is very difficult to measure the size of the external benefits, and therefore to calculate precisely which goods and services should be supported and the level of support they should receive. In addition, both direct provision and subsidies are often highly political in nature, as different groups compete with each other over who will receive the most benefits. Governments are often susceptible to political pressures and sometimes make choices based on political rather than economic criteria.

(HL only) Nudges are subject to the same limitations that were noted in connection with negative consumption externalities (Chapter 5), involving difficulties in designing effective nudges in view of insufficient knowledge about how people respond to nudges and choice architecture, as well as the possible different responses these may have across income and cultural groups.

Therefore, in the real world it is very unlikely that governments are able to shift the MPC or MPB curves by the amount necessary to correct the positive externalities. The most that can be hoped for is that the policies in question will be a step in the right direction.

Legislation, education and awareness creation are subject to similar limitations concerning calculating the size of external benefits. Only sometimes can they be effective, and then can only help shift the MPB curve in the right direction, rather than achieve a demand increase that will bring the economy to the Q_{opt} level of output. For example, they can have very positive effects in certain cases (such as legislation requiring schooling up to a minimum age or education on the importance of good nutrition), but in other cases are ineffective (for example, they cannot on their own increase consumption of health care services and education to the optimum level). Moreover, they have the further effect of raising the price of the good to consumers, which may make the good unaffordable for some consumer groups. Therefore, legislation and awareness creation sometimes can be used more effectively if they are implemented together with direct provision and subsidies. A good example is education, where compulsory schooling up to a certain age (legislation) goes together with direct government provision.

TEST YOUR UNDERSTANDING 6.2

1 a Using a diagram, show how marginal private benefits and marginal social benefits differ when there is a positive consumption externality.

b Explain the difference between the equilibrium quantity determined by the market and the quantity that is optimal from the point of view of society's preferences.

c Describe the problem with the allocation of resources achieved by the market when there is a positive consumption externality.

d Show the welfare loss created by the positive consumption externality in your diagram, and explain what this means.

CONTINUED

2 Provide some examples of positive consumption externalities.

3 For each of the examples you provided in question 2, explain some methods that could be used to correct the externality.

4 Outline how a positive consumption externality differs from a positive production externality.

5 Explain the meaning of a merit good, and provide examples.

6 Discuss advantages and disadvantages of the policy measures that governments can use to correct positive externalities of production and consumption.

REAL WORLD FOCUS 6.1

Positive externalities of cash transfer programmes

A highly successful policy used in Brazil to reduce poverty and income inequalities is the Bolsa Familia programme implemented by the Brazilian government. Through this programme poor families receive cash transfers *on condition* that they send their children to school and get them vaccinated. These programmes are therefore called *conditional cash transfers* (CCTs). The programme is so successful in reducing poverty over the short term through the cash transfers and over the long term by increasing human capital that it has attracted great attention around the world.

Figure 6.7: Feijao, Acre Province, Brazil. School children in a rural school classroom

A study on crime in Brazil has found that the Bolsa Familia programme has had a significant impact on crime reduction in São Paolo. It concludes that in the areas where CCTs are implemented there is a negative impact on crime. This is not related to the increased time spent in school but rather is due to reduced poverty and inequality, resulting in a reduction primarily of robberies, hence economically motivated crimes.

Source: *Spillovers from conditional cash transfer programmes http://ftp.iza.org/dp6371.pdf IZA, DP No. 6371*

Applying your skills

You may note that this is a type of positive externality that cannot be analysed by means of the usual externality diagram as there is no market that can be easily identified here. However the idea of a positive externality resulting from government spending on CCTs is clear. Explain what kind of externality this represents, identify the external benefits and describe the implications for the government's CCT policy.

Summary of externalities

Table 6.1 summarises important information on each type of externality.

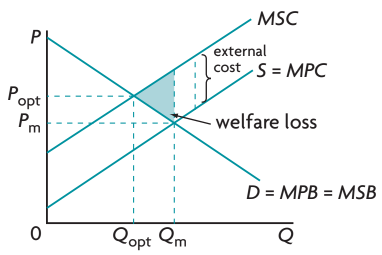
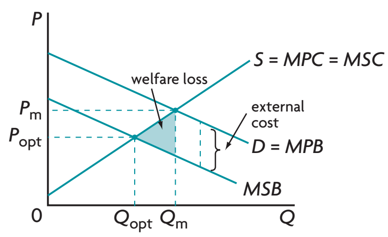
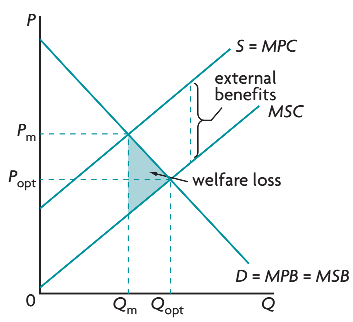
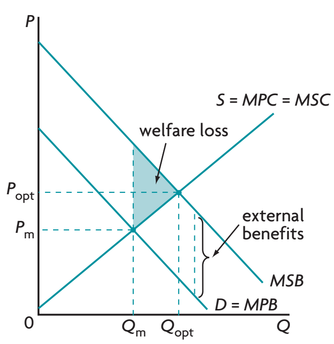
Type of externality	Examples	Policies
Negative production and many common pool resources  	Producers impose external costs on society Production by use of fossil fuels; external costs include global warming, negative effects on health, environmental pollution	• Indirect (Pigouvian) taxes • Carbon taxes • Tradable permits • Legislation, regulation • Collective self-governance • Education, awareness creation • International agreements
Negative consumption  	Consumers impose external costs on society Use of cars and heating using fossil fuels; external costs include global warming, negative effects on health, environmental pollution	• Indirect (Pigouvian) taxes • Legislation, regulation • Education, awareness creation • Nudges (HL only)
Positive production  	Producers create external benefits for society Research by private firms leads to development of new technologies that benefit the whole of society	• Government provision • Subsidies
Positive consumption  	Consumers create external benefits for society Education and health care lead to benefits for the whole of society, including lower unemployment, lower crime rates, higher economic growth	• Legislation, regulation • Education, awareness creation • Nudges (HLonly) • Government provision • Subsidies

Table 6.1: Summary of externalities

The information in this section applies to *all externalities*, including the negative ones presented in Chapter 5 and the positive ones of this chapter.

How to draw an externality diagram without memorising

Students often have difficulty learning how to draw the externality diagrams and labelling the curves correctly. The following rules will help you draw any externality diagram without memorising.

The rules

1. In a production externality, the supply curve splits into two; in a consumption externality, the demand curve splits into two.
2. Supply reflects costs; demand reflects benefits.
3. The market equilibrium quantity Q_m corresponds to private costs and benefits, *MPC* and *MPB*; the social optimum reflects social costs or benefits.
4. In a negative externality $Q_m > Q_{opt}$, meaning that *the market provides too much of a bad thing*; in a positive externality $Q_m < Q_{opt}$, meaning that *the market provides too little of a good thing*.

How to use the rules

1. Draw a demand and supply diagram and label the axes P and Q.
2. For a production externality draw two parallel upward sloping curves; for a consumption externality draw two parallel downward sloping curves. Find the two equilibrium quantities on the Q axis (but do not label them yet).
3. In a negative externality, since $Q_m > Q_{opt}$, label the larger quantity Q_m and the smaller quantity Q_{opt}; in a positive externality, since $Q_m < Q_{opt}$, label the larger quantity Q_{opt} and the smaller quantity Q_m.
4. Using rule 3 above, Q_m gives *MPC* and *MPB*, while Q_{opt} gives *MSC* and *MSB*. You can now label all the curves. (Note that the demand curve, D and supply curve, S represent *private* benefits and costs; therefore D = *MPB* and S = *MPC*.)
5. Find the triangle that points to Q_{opt}, and that lies in between the two curves that have split. This is welfare loss.

Some useful information to remember about externalities

Here are some points to bear in mind:

- All negative externalities (of production and consumption) *create external costs. When there are external costs, MSC > MSB at the point of production by the market.*
- All positive externalities (of production and consumption) *create external benefits. When there are external benefits MSB > MSC at the point of production by the market.*
- All production externalities (positive and negative) create *a divergence between private and social costs (MPC and MSC)*.
- All consumption externalities (positive and negative) create *a divergence between private and social benefits (MPB and MSB)*.

6.3 Market failure and public goods

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain public goods as being (AO2)
  - non-rivalrous, non-excludable
  - subject to the free-rider problem
- evaluate government intervention to provide public goods (AO3)
  - direct provision
  - contracting out to the private sector

The meaning of public goods

Public goods are non-rivalrous and non-excludable

We learned the meaning of rivalry and excludability in Chapter 5. A good is rivalrous when its consumption by one person reduces its availability for someone else. It is excludable if it is possible to exclude people from using it. Private goods as we have seen are rivalrous and excludable. Common pool resources are rivalrous and non-excludable. Public goods, by contrast, are non-rivalrous and non-excludable.

A **public good** has the following two characteristics:

- It is **non-rivalrous**; its consumption by one person does not reduce consumption by someone else.
- It is **non-excludable**; it is not possible to exclude someone from using the good.

For example, a lighthouse is non-rivalrous, because its use by one person does not make it less available for use by others. Also, it is non-excludable, because there is no way to exclude anyone from using it. Other examples of public goods include the police force, national defence, flood control, non-toll roads, fire protection, basic research, anti-poverty programmes and many others.

Public goods and the free rider problem

How do public goods relate to market failure? In the case of excludable goods, it is possible to prevent people from buying and using a good simply by charging a price for it, since those who are unable or unwilling to pay the price do not get to use it. Therefore, private firms have an incentive to provide excludable goods because they can charge a price for them, and therefore can cover their costs. Non-excludable goods differ: if a non-excludable good were produced by a private firm, people could not be prevented from using it even though they would not pay for it. Yet no profit-maximising firm would be willing to produce a good it cannot sell at some price. As a result, the market fails to produce goods that are non-excludable, giving rise to resource misallocation, as no resources are allocated to the production of public goods.

Public goods illustrate the **free rider problem**, occurring when people can enjoy the use of a good without paying for it. The free rider problem arises from non-excludability: people cannot be excluded from using the good. Public goods are a type of market failure because, due to the free rider problem, private firms do not produce these goods: the market fails to allocate resources to their production.

Table 6.2 summarises different types of goods according to their rivalry and excludability characteristics.

	Rivalrous	Non-rivalrous
Excludable	Private goods Goods with or without positive or negative externalities (both production and consumption) sold for a price. Merit goods (as long as they are produced by the market) and demerit goods Examples: computers, books, clothes, education, petrol (gasoline)	Quasi-public goods Goods that do not fall neatly into the other three categories; often (but not always) have large positive externalities so may be provided by the government Examples: uncrowded toll roads, museums, public swimming pools that charge entrance fees, cable TV See the section below for an explanation of quasi-public goods.
Non-excludable	Common pool resources Natural resources that are not owned by anyone, not sold in markets and not having a price; their lack of a price makes them subject to overuse (unsustainable use), depletion and degradation Examples: forests, rivers, lakes, soil quality, fish in the oceans	Public goods See the section below for an explanation of quasi-public goods. Socially desirable goods not produced by private firms because it is not possible to charge a price; they are subject to the *free rider problem*: people use them without having to pay; since they are socially desirable they are produced by the government and provided free of charge Examples: national defence, street lighting

Table 6.2: Summary of different kinds of goods based on rivalry and excludability

Quasi-public goods (Supplementary material)

Some goods do not fit neatly into the category of private goods or public goods. They can be considered to be 'impure' public goods, also known as 'quasi-public goods'. These goods are:

- non-rivalrous (like public goods), and
- excludable (like private goods).

Examples include public museums that charge an entrance fee and toll roads. All these are excludable because consumers must pay to use them. They are also non-rivalrous since use by one does not reduce availability for others. Since the price system can be made to work here to exclude potential users, they could be provided by private firms. However, they all have very large positive externalities, thus justifying direct government provision.

Government intervention to correct the market's failure to provide public goods

Direct government provision

We have seen that the market fails to allocate resources to the production of public goods. This means the government must step in to ensure that public goods are produced at socially desirable levels. Thus public goods are directly provided by the government, are financed out of tax revenues and are made available to the public free of charge (or nearly free of charge).

Government provision of public goods raises some issues of choice about (a) which public goods should be provided, and (b) in what quantities they should be provided. These issues are similar to what was noted earlier in connection with direct government provision and subsidies for goods with positive externalities. Limited government funds force choices on what public goods to produce, and each choice has an opportunity cost in terms of other goods and services that are foregone (sacrificed).

Here, too, the government must use economic criteria to decide which public goods will provide the greatest social benefits for a given amount of money. However, in the case of public goods, governments face a major additional difficulty in calculating expected benefits. With private goods that are provided or subsidised by the government, it is possible to make estimates of expected benefits by using the market price of the good. (Remember the market price reflects the benefits consumers receive and so reveals its value to consumers.) Therefore, the government can use the market price of private goods with positive externalities to estimate their value to consumers, but with public goods there is no such possibility as they are not produced by private firms and have no price.

This means the government must try to estimate the demand (or 'price') of public goods through such means as votes or surveys of people who are asked how much a good would be worth to them. This information is used in *cost–benefit analysis*, which compares the estimated benefits to society of a particular good with its costs. If the total benefits expected from a public good are greater than the total costs of providing it, then the good should be provided. If benefits are less than costs, then it should not be provided. Assuming that cost–benefit analysis indicates a public good should be provided, the decision on how much of it to provide is made by comparing marginal benefits with marginal costs: the public good should be provided up to the point where $MB = MC$.

Whereas the costs of providing a public good are relatively easy to estimate, there are clear difficulties in estimating benefits. A major difficulty arising with surveys is that people who really want something are likely to exaggerate its value. Therefore, cost–benefit analysis is a very rough and approximate method used to make choices about public goods.

In addition, it should be noted that just like in the case of positive externalities which invite direct government provision and subsidies for their correction, the provision of public goods is also political in nature, with different groups competing against each other, and with government sometimes susceptible to political pressures that influence their decisions.

Contracting out to the private sector

When the government provides a public good, it may either provide it itself directly, or it may contract it out to a private firm. **Contracting out** by the public sector to the private sector occurs when a government makes an agreement (or contract) with a private firm to carry out an activity that the government was previously doing itself. This is a practice that many governments around the world have been increasingly undertaking. Suppose the government would like to build a new highway system. Rather than build it itself by directly hiring engineers, workers, and supplying the materials, it may contract its construction out to a private construction firm.

Note that whether goods are provided directly by the government or contracted out, they are in both cases financed out of tax revenues.

The potential advantages of contracting out include the following:

- It is often done by competitive tendering (a competition between firms that would like to provide the government with the service) resulting in the selection of a provider that can offer the lowest cost.
- It is usually accompanied by detailed specifications regarding the activity that will be provided with criteria for measurement of the provider's performance, which may allow for better quality control.
- It provides access to a broader range of skills and technology of the private firm than the government is likely to have available itself.
- The private firm may be more flexible and innovative than the government.
- Due to all of the above it may be possible for the public goods provided to be better quality and less costly.

On the other hand, contracting out also has potential disadvantages:

- The government becomes less accountable for the public goods it provides.
- The government loses control over the services it has contracted out.
- The costs of contracting out may be greater than if the government had provided the public good itself, as the contracting private firms often charge high prices for their services.
- There is a risk that quality may be reduced because competition between firms on the basis of cost may lower the quality of the services provided.
- There is a risk of making a poor contract with a private sector firm, resulting in higher costs and lower quality, along with reduced accountability and control noted above.
- Contracting out needs to be monitored by the government, which adds to costs.

As a result of all of the above it is difficult to generalise about the effectiveness of contracting out as a method to provide public goods.

Moreover, it should be noted that contracting out only addresses issues of the quality, costs, skills, and so on discussed above; it does not address the difficulties faced by the government in making decisions about what public goods to produce and in what quantities, which as we have seen above depend on the very difficult problem of trying to determine expected benefits and costs of alternative public goods.

TEST YOUR UNDERSTANDING 6.3

1. Explain how the concepts of rivalry and excludability relate to the distinction between public goods and private goods.
2. Provide some examples of public goods, and outline how they relate to the concepts of non-rivalry and non-excludability.
3. Use the concepts of resource allocation and the free rider problem to explain how public goods are a type of market failure.
4. Discuss the difficulties of direct government provision of public goods.
5. Evaluate the government's policy option to contract out to the private sector the provision of public goods.

6.4 Asymmetric information (HL only)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between adverse selection and moral hazard (AO2)
- evaluate government responses to the problem of asymmetric information, including legislation and regulation, and provision of information (AO3)
- evaluate private responses to the problem of asymmetric information, including signalling and screening (AO3)

The competitive market mechanism presupposes that all firms and all consumers have complete information regarding products, prices, resources and methods of production. However, as we know from Chapter 2 this condition is hardly ever met in the real world where firms, consumers and resource owners find themselves in situations where information is missing. **Asymmetric information** refers to something more than just missing information; it refers to situations where buyers and sellers do not have *equal access* to information. In some cases buyers may have more information than sellers; in other cases the opposite holds where sellers have more information than buyers. As we will see, this usually results in an *underallocation of resources* to the production of goods or services, and therefore to allocative inefficiency.

We will study two different types of problems of asymmetric information: adverse selection and moral hazard.

Adverse selection

Adverse selection refers to situations where one party in a transaction has more information about the quality of the product being sold than the other party.

Adverse selection where information is available to sellers but not to buyers

Sellers often have information about the quality of a good or service that they do not make available to consumers. Sellers of used cars have information about the car's quality that they are unlikely to reveal to potential buyers if the car has a defect. In a free and unregulated market, sellers of food could sell products that are unsafe for human consumption, possibly leading to illness and even death. Sellers of medicines could sell unsafe medications that could be ineffective or dangerous. Individuals claiming to be doctors, some of whom have little or no training, could practise medicine and even surgery, resulting in huge costs in terms of human health and safety.

In a free unregulated market, the result is usually to underallocate resources to the production of the good or service. Consumers are likely to be aware of possible dangers to themselves, and will be cautious about buying the good or service, resulting in lower demand, less production and lower sales. However, if consumers are unaware of possible hidden dangers, such as with unsafe food or toys, there could result an overallocation of resources to the production of these goods and services.

Possible solutions to adverse selection where the seller knows more than the buyer

Government responses

Regulation

According to one method, governments can pass laws and regulations that ensure quality standards and safety features that must be maintained by producers and sellers of goods and services, such as food, medications, private schools, toys, buildings and all types of construction.

These methods are not without their difficulties. Legislation and regulation are time-consuming, bureaucratic procedures, which sometimes work to slow down economic activities. It takes a long time, for example, to test new medications and certify their safety, and is a costly process. Also, regulatory and quality control activities have very large opportunity costs. Just in the case of food safety control, which involves not only food and beverage products but also hygiene in restaurants, there is a huge number of products and service providers involved, who require regulation and monitoring from the level of the farm (regarding the kinds and amounts of chemical inputs) to the moment the food reaches the table.

Provision of information

Governments may also respond by directly supplying information to consumers, or by forcing producers to provide information, thus protecting consumers in their purchasing decisions. This may include information about the quality of medical care by different providers, about communicable diseases, crime rates by neighbourhoods, health hazards related to different activities, products or substances, nutritional labelling on foods, and so on. In some countries, particularly in Europe, governments provide fee schedules for services (such as legal, medical, architectural) to ensure that consumers receive a particular quality of services for a particular range of prices.

When the government is the provider of information, there are difficulties involving the collection and dissemination of all the necessary information to consumers, the accuracy of the information, as well as opportunity costs in providing the information. When a private seller/producer is the provider of the information, there are serious questions about whether information regarding all hazards in products or substances and materials used in products, or all information regarding the quality of services (whether legal, medical, financial, and so on) is accurate and complete.

Another problem is that it is sometimes not possible to eliminate an information asymmetry between sellers and buyers, because no matter what regulations and information are provided, there is still some room for the seller to hide some information from the buyer. In the areas of health care and law, doctors and lawyers have specialised, technical information about their clients that the clients themselves do not possess. Doctors and lawyers often use this information for their own private gain by selectively revealing information to their clients that causes them to demand more services than are necessary. This practice leads to what is known as 'supplier-induced demand', or demand that is induced (created) by the supplier, which would not have appeared if the client had equal access to information.

Licensure

In the case of doctors, most countries around the world have laws requiring doctors to be licensed, and a licence can only be obtained upon proof of adequate medical competence. Licensing is similarly required for many other professions in many countries, from teachers and lawyers to plumbers and electricians.

Some economists criticise licensing, because it may work to limit the supply of people in a profession, raising the price of their services and increasing their incomes at the expense of consumers who must pay higher prices (this refers to *market power* discussed in Chapter 7 at HL).

Private responses

Screening

Screening is a method used by the party with the limited information, in this case the buyer. The buyers may try to get more information about what they are buying, in other words they screen the product or the producer or seller of the product. Consumers may find information provided on the internet about reliable used car dealers, or they may informally ask friends for information about the quality of health care service providers, or legal service providers.

While screening may help in providing consumers with some missing information, it can by no means on its own solve the problem as it cannot provide systematic and complete information to match the information available on the seller's side.

Signalling

Signalling is a method used by the party that has more information, or in this case the seller. The purpose of signalling is to convince the buyers that the product being sold is of good quality. Common methods include the use of warranties, establishment of brand names that convey a feeling of reliability, and, in the case of used cars, making service records available or exhibiting cars in fancy show rooms.

The problem with signalling is that it is unlikely to provide full information to buyers, and it may even provide inaccurate or misleading information by sellers eager to promote and sell their products.

Adverse selection where information is available to buyers but not to sellers

Adverse selection where the buyer has information not available to the seller often arises in the area of insurance services, where the buyer of insurance has more information than the seller. This situation arises most often in the area of health insurance. Buyers of health insurance know more about the state of their health than sellers of insurance, and those with health problems are unlikely to tell the full truth to the insurance company. In a free unregulated market, adverse selection results in an underallocation of resources to health insurance services, as the insurance company reduces the supply of insurance to protect itself against having to provide insurance coverage to very high risks, or people who are more likely to become ill. Adverse selection also leads to high insurance costs for insurance buyers.

Possible solutions to adverse selection where the buyer knows more

Private responses

Private insurance companies usually protect themselves by offering a range of policies where the lower the cost of the insurance, the higher the deductible (out-of-pocket payments). This offers people choice, so that those who believe they have a low risk of getting sick can buy a low-cost policy with a higher deductible, while higher-cost policies with lower deductibles can be selected by people who believe they have high levels of health risk. This is actually a method of *screening* undertaken by the party with the limited information, or seller of insurance. The choice of high or low deductible given to the buyers of insurance is intended to screen them by indirectly providing information about their state of health to the seller of insurance.

However in practice, it does not work out this way. An important reason is that lower-income earners choose low-cost policies with high out-of-pocket payments because these are more affordable, *regardless of the* *state of their health*. From the perspective of equity or fairness, this is undesirable because it discriminates against those on low incomes. Another reason is that in trying to protect themselves against high risks, insurance companies usually refuse to insure people above a certain age, as elderly people generally have a higher chance of becoming ill. The result is that those who mostly need health insurance coverage, who are poor people who cannot afford to buy health care in the private market and elderly people who are more likely to become ill, are left with little or no insurance coverage.

Government responses

To deal with this problem, government responses may take the form of direct provision of health care services at low or zero prices to an entire population, financed by tax revenues, thus ensuring that the entire population has health insurance coverage, such as in countries with a National Health Service (as in many European countries). Alternatively, they make take the form of social health insurance, which may cover a country's entire population (as in several European countries), or which selectively covers only certain vulnerable groups of the population (as in the United States). The benefit of these approaches, particularly in countries that offer insurance coverage to their entire population, is that no one in need of health care goes without it.

A potential problem with government-funded or social health care systems involves difficulties in controlling costs of providing health care and growing burdens on the government budget or social health insurance budget.

Moral hazard

Explaining moral hazard

Moral hazard refers to situations where one party takes risks, but does not face the full costs of these risks because the full costs of the risks are borne by the other party. It usually arises when the buyer of insurance changes his or her behaviour *after* obtaining insurance, so that the outcome works against the interests of the seller of insurance. For example, buyers of car theft insurance may be less careful about protecting their car against theft, because they know they will be reimbursed if someone steals their car. Some buyers of medical malpractice insurance (doctors) may be less careful about avoiding malpractice, because of the knowledge that malpractice costs will be covered by the insurer. Unemployment insurance may lead some people to be less hesitant about becoming unemployed, in the knowledge that their insurance will provide them with some income.

In all these cases, the buyers of insurance have information about their future intentions that is not available to the sellers of insurance. In a free, unregulated market, the result of moral hazard is to underallocate resources to the production of insurance services, as sellers of insurance try to protect themselves against higher costs due to the risky behaviour of the buyers of insurance.

Many economists have noted that the financial crisis that began in 2008 was partly a case of moral hazard. It is argued that many financial institutions made risky loans and engaged in other highly risky financial transactions because they believed that the government would support them in the event of difficulties (which, in fact, it did). (Note that this is not a matter of taking out an insurance policy in the strict sense of the term, but it is a sort of 'insurance' nonetheless, in that the government provides a kind of assurance of protection in the event that financial institutions face difficulties due to poor loan repayments.)

You may note that the term 'moral hazard' does not refer to unethical or immoral behaviour. It is simply a historical remnant of a very old insurance term that originally meant 'subjective'.

Responses to moral hazard

Problems of moral hazard in insurance are usually dealt with by the provider of insurance. This is often done by making the buyer of insurance pay for part of the cost of damages through deductibles (out-of-pocket payments). This is intended to make the insurance buyer face the consequences of risky behaviour, thus leading to less risky behaviour. It will be recalled from the discussion above that deductibles are a form of *screening*.

A problem with deductibles is that it has different effects depending on the income level of insurance buyers. As noted earlier in connection with adverse selection, private insurance companies usually offer a range of policies from which buyers can choose, where the lower the cost of the insurance, the higher the deductibles. Higher-income earners usually choose higher-cost policies with low deductibles, while lower-income earners choose low-cost policies with high deductibles because these are more affordable. This suggests that higher-income earners are more likely to engage in risky behaviour because they are offered more insurance protection, while lower-income earners are less likely to engage in risky behaviour.

In the financial area, moral hazard is dealt with through government regulation of financial institutions, intended to oversee and prevent highly risky behaviour. This raises a whole set of issues regarding the types and degrees of government regulations that are required if these are to be effective. In general, following the onset of the global financial crisis in 2008 governments in Europe and the United States have taken steps to increase regulation of the financial sector, though there are concerns that this has not been enough, especially in the United States.

TEST YOUR UNDERSTANDING 6.4

1 a Using examples, identify two main types of information problems that give rise to market failure.

b Using the concept of allocative efficiency, explain why information asymmetries represent market failure.

2 a Provide examples of information asymmetries where information is available to sellers but not to buyers.

b Explain how governments can intervene to correct these information asymmetries.

c Explain possible private responses to information asymmetries.

d Discuss some advantages and disadvantages of both government intervention and private responses.

3 a Provide examples of information asymmetries where information is available to buyers but not to sellers.

b Explain how governments can intervene to correct these information asymmetries.

c Discuss advantages and disadvantages of these types of government intervention.

6.5 Equity in the distribution of income and wealth (HL only)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain why the free market results in an unequal distribution of income and wealth (AO2)
- draw a circular flow of income diagram to show why free markets give rise to income inequalities (AO4)

Why the free market economy results in income inequalities

In Chapter 1, we saw that *equity* refers to the idea of being fair and just, while *equality* is the state of being equal with respect to something. Therefore *income equality* would mean that everyone in a society receives the same amount of income. We also learned that while equity and equality have different meanings, in most countries the pursuit of equity is understood to refer to efforts to reduce significant inequalities in income and wealth. The reason for this is that people around the world generally share the belief or value judgement that the free market economy results in inequalities that are considered to be unfair. We will now see why this is so.

Our study of the circular flow model in Chapter 1 shows that in a market economy the amount of goods and services that households receive depends on their income, as this determines how much they can buy. Figure 1.4, showing the simple circular flow model (with no leakages and injections), appears below as Figure 6.8. We can see here that the income of households depends on payments they receive by selling the factors of production they own. Therefore, output and income distribution in a market economy depend on how many resources consumers (households) own and are able to sell in resource markets, as well as on the prices of the factors of production they sell.

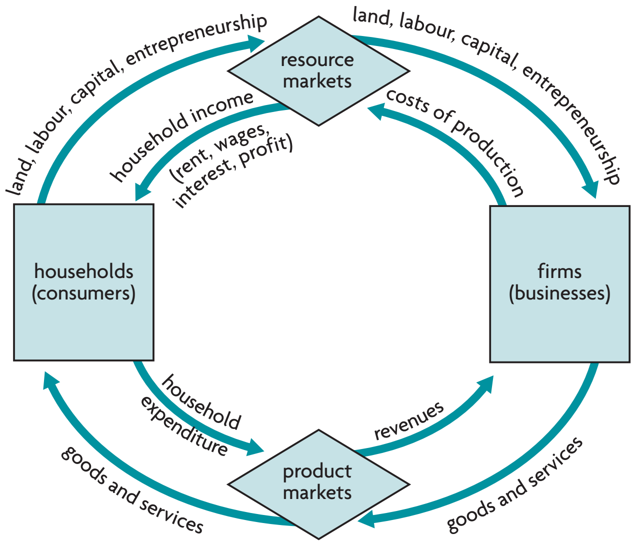


Figure 6.8: Circular flow of income model

The problem of income distribution arises because *ownership of factors of production is highly unequal, and because the prices of factors of production determined in the market vary enormously*. Most people have labour resources that they provide in labour markets, for which they receive wages. Yet some people are able to receive very high wages because of special skills and education or natural talents, while others who are less skilled, educated or talented may receive wages so low that they may be unable to cover the most basic needs for themselves and their families (food, shelter, clothing, etc.).

Further, there may be people who would like to work but cannot do so because they lack the kinds of skills firms want to hire, or because they are sick, or old, or have special needs that prevent them from working, or because there simply are not enough jobs in the economy to provide work for everyone. Then there are some individuals who possess some of the additional factors of production of land, capital and entrepreneurial abilities, for which they receive rental, interest and profit income. These additional factors of production are generally highly unequally distributed. As a result, the market-determined distribution of income in an economy is likely to be very unequal, with some people receiving far larger shares than others.

It follows that markets cannot ensure that everyone in a population will secure enough income to satisfy their basic needs. Most societies consider this to be a disadvantage of the free market economy, because of the belief held by most people that everyone in a population should be able to satisfy a minimum of basic human needs. Governments around the world therefore use a variety of methods to change the market-determined distribution of income and output, and arrive at a more socially desirable outcome. This is referred to as *redistribution*. We will come back to the topic of income distribution in Chapters 12 and 20.

Inequalities in wealth

As we have seen in connection with the circular flow model, income refers to the flow of money that is received by the owners of the factors of production. **Wealth**, on the other hand, refers to the money or things of value that people own, such as savings deposits (money saved in a bank); stocks in the stock market; bonds; land, houses and other property; valuable paintings or jewellery, and so on; minus debt to banks or other financial institutions.

Income gives rise to the possibility of saving, which can then be used to create wealth, and so wealth and income are related to each other as would be expected. Yet they are distinct and are not always closely linked together. For example, some people may have high incomes but low savings because they spend a lot, in which case they will have relatively low wealth. On the other hand, there may be people who have inherited wealth, or pensioners who have saved over a lifetime of work, who have high levels of wealth but relatively low incomes.

In general, however, we can say that the higher the income, the greater the possibilities for saving and for accumulating wealth. Therefore, just as the free market economy results in income inequalities, so too it results in wealth inequalities.

We will come back to the topic of inequalities in income and wealth in Chapter 12, where we will discover that *inequalities in wealth are in fact far greater than inequalities in income*.

Inequality in income and wealth in relation to market failure

It should be noted that the inability of the market to ensure that everyone in a population will secure enough income to satisfy their basic needs *is not strictly speaking market failure*. This follows from the way that market failure is defined. As we know, market failure is the inability of the market to achieve allocative efficiency, where $MSB = MSC$, or where social surplus is maximum. This bears no connection with income or wealth distribution. It is in fact possible to have allocative efficiency and maximum social surplus either with complete income equality or with extreme inequality, where for example a large portion of the population is starving. This peculiarity will be discussed in the Theory of knowledge 6.1.

TEST YOUR UNDERSTANDING 6.5

1 Explain, using examples, the difference between equity and equality in income wealth distribution.

2 Using the circular flow of income diagram, explain why the market system cannot ensure that everyone in a society will be able to secure an adequate income to satisfy basic needs.

THEORY OF KNOWLEDGE 6.1

Inequality in relation to market failure: the possibility of maximum social surplus in the presence of extreme inequalities

It was noted above that inequalities in income and wealth are not a type of market failure, because it is possible to have any degree of inequality where there is at the same time allocative efficiency with maximum social surplus.

Yet you may wonder, how can there be maximum social surplus with extreme inequality? The answer to this lies partly in the positive-normative distinction we discussed in Chapter 2. If you revisit Theory of knowledge 2.1 in Chapter 2 (at the end of Section 2.5), you will be reminded that maximum social surplus (or welfare) refers to the *what/how much to produce* and *how to produce* questions of economics, and the idea of making the best possible use of resources, which are in the sphere of positive economics. Equality and inequality on the other hand refer to the *for whom to produce* question, which belongs to the sphere of normative economics.

Yet this by itself is not a satisfactory explanation, especially if we remember that allocative efficiency can be defined as *producing the combination of goods mostly wanted by society*. In Chapter 2, Section 2.5, we saw that allocative efficiency is achieved when the economy allocates its resources *so that the benefits from consumption are maximised for the whole of society*.

This clearly gives rise to a puzzle. How can benefits from consumption be maximised for the whole of society if there is extreme inequality with a large part of the population starving?

We can find the answer to this puzzle in the definition of *demand*, which we studied in Chapter 2 (Section 2.1). As you may recall, demand shows the various quantities of a good that a consumer is willing and *able to buy* at different possible prices, *ceteris paribus*. The answer to our puzzle lies in the *able to buy* part of this definition. If consumers have little or no income, they are not able to buy the good, and therefore they do not have any demand for it. In other words, no demand will show up in the market for the good from consumers with little or no income, however much they may want to buy it.

We now have the answer to the puzzle. As we know, the condition $MSB = MSC$ is equivalent to maximum social surplus (or welfare). Alternatively, we can say that $MB = MC$ is equivalent to maximum social surplus when there are no externalities. But MB is simply the demand curve! And MSB is also simply a demand curve for the whole of society that accounts for any possible external costs or benefits! This means that when we say that there is allocative efficiency and maximum social surplus when *the benefits from consumption are maximised for the whole of society*, we are simply talking about that part of society that has a demand because consumers are not only willing but also *able to buy* the good! Consumers from the rest of society with little or no income and hence no demand are bypassed, ignored and forgotten.

CONTINUED

In other words, allocative efficiency is defined on the basis of demand from those consumers who have enough income to make their preferences felt in the market.

Thinking points

The work of the classical economists, such as Adam Smith who we met in Chapter 1, was not based on the positive-normative distinction. Economists at the time, as you may remember were moral philosophers, and Adam Smith was deeply concerned with issues of ethics. It was only in the early 20th century that economists became influenced by the methods of the natural sciences and the idea that *facts* should be kept distinct from *values*. Yet in more recent years, some economists, such as Amartya Sen, an Indian Nobel Prize winning economist for his work on Economic Development, argue that it does not make sense for economists to maintain this strict separation between facts and values. Sen's work reintroduces ethics and values into economic analysis and mixes them to create a unified whole. We will encounter Amartya Sen in Chapter 18.

- In your study of economics, you are encouraged to consider the positive-normative distinction. Do you agree with the idea that facts (positive) should be kept distinct from values (normative) in the study of economics? Is it even possible to always make this distinction, or do values creep into the work of economists? We will come back to this point in Theory of knowledge features 9.1, 10.1, 15.1 and 18.1.
- Consider what redistribution of income leading to reduced inequalities in an economy implies for the concept of allocative efficiency. In particular, consider that people who previously had no demand for a good or service may now have such a demand based on their higher income. What is likely to happen to the efficient level of output of particular goods and services?

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1. Identify a good or service in your country or a country you are interested in that gives rise to a particular positive consumption externality. Examine the external benefits and their effects on stakeholders, investigate solutions in use, evaluate the solutions and identify effects on stakeholders.
2. Identify a public good that has been contracted out to the private sector in a country you are interested in and investigate the advantages and disadvantages that have emerged.
3. Investigate a situation where asymmetric information involves adverse selection or moral hazard and identify public and private solutions that have been used to address it.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 7

Market failure and socially undesirably outcomes III:

Market power (HL only)

BEFORE YOU START

- In what ways do firms compete with each other in order to increase their sales?
- When firms compete with each other to increase sales, some firms might have advantages over others. Can you think of what some of these advantages may be?
- From the consumer's point of view can you think of some advantages and disadvantages of firms that grow to a very large size?

This chapter is concerned with the relationship between the behaviour of firms and market failure. We will begin by introducing the fundamental concepts of revenues, costs and profits needed to study firm behaviour. We will then see how firm behaviour within particular markets gives rise to market power, which is linked to an important kind of market failure.

Photo: London, England. Piccadilly at night

7.1 Introduction to firms, industries and market structures

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the main characteristics of perfect competition – many firms, free entry, homogeneous products (AO2)
- explain the main characteristics of monopoly – single or dominant firm, high barriers to entry, no close substitutes (AO2)
- explain the main characteristics of market structures in imperfect competition (AO2)
    - varying degrees of market power, firm is price maker
    - monopolistic competition – many firms, free entry, product differentiation
    - oligopoly – few large firms, high barriers to entry, interdependence
- explain the meaning of market power (AO2)

Market power and market failure

As you know from past chapters, a *firm* (or business) is an organisation that employs factors of production to produce and sell a good or service. A group of one or more firms producing identical or similar goods or services is an *industry*. The car industry consists of car manufacturers (Ford, Honda, Mercedes, etc.); the shoe industry consists of shoe manufacturers; the banking industry consists of providers of banking services (banks). Industries are analysed by use of models called *market structures,* which describe characteristics of a market organisation that determine the behaviour of firms.

A fundamental distinction between market structures is *the extent to which each individual firm in the industry is able to control the price at which it sells its product*. This is known as **market power**.

On the basis of this idea, we can distinguish between:

- the extreme of **perfect competition**, where firms have no ability to control the price of their products; such firms have *zero market power*
- the extreme of **monopoly**, where there is single firm in the market; this firm has the greatest ability to control the price of its product, and therefore *the greatest amount of market power*.

In between the two extremes are the two other market structures where *firms face some competition but also have market power*, and for this reason they are known as **imperfect competition**; they include:

- **monopolistic competition**
- **oligopoly**.

Market power is important because it gives rise to market failure, or the failure to achieve allocative efficiency. The greater the market power, the greater the allocative inefficiency.

Firms in perfect competition, which have zero market power, do not result in market failure, and are the only market structure where firms achieve allocative efficiency. Firms in the other three market structures result in market failure to varying degrees, with monopoly resulting in the greatest degree of allocative inefficiency.

Introduction to market structures

Apart from differences in their ability to control price (market power), market structures can be defined on the basis of three main characteristics:

- the *number of firms* in the industry
- how similar or different are the goods or services produced, known as **product differentiation**, meaning that the products can be made different from each other, or not identical
- how easy or difficult it is for new firms to enter the industry and begin producing, which depends on whether there are **barriers to entry**; these include anything that can prevent a firm from entering an industry and beginning production. When there are no barriers to entry into an industry we say there is **free entry**.

Characteristics of perfect competition

- There is a very large number of firms in the industry.
- All the firms in the industry sell **homogeneous** products; this means the products are identical (undifferentiated); there are no brand names
- There are no barriers to entry; any firm that wants to enter the industry and begin producing the good or service can do so freely.

Perfect competition is highly unrealistic and it is difficult to find markets in the real world that meet the characteristics above. Examples that come the closest include the international markets for agricultural commodities such as wheat, rice, corn and livestock, other commodities like silver and gold, stock and bond markets, and the foreign exchange market (in which currencies of different countries are bought and sold). In these cases, each firm is small relative to the total number of firms so that each one does not have much ability to influence its price. Further, the product sold by all firms is identical or nearly identical, and any new firm is free to enter the industry.

Characteristics of monopoly

- There is a single seller or dominant firm in the industry, so the single firm is the entire industry. In the real world, a monopolistic industry may consist of one firm that dominates the market with a very large market share. For example, Monsanto controls 80% of the genetically modified corn market and is considered to be a monopoly.
- The firm produces and sells a unique good or service, with no close substitutes; if substitute goods existed, then consumers could switch to buying the substitute, and there would no longer be a monopoly for the good.
- There are high barriers to entry in the industry; the monopolist owes its dominance in the market partly to the inability of other firms to enter.

Examples of monopolies include telephone, water and electricity companies in areas where they operate as a single supplier. (These are known as natural monopolies to be discussed below.) These may be in local areas in which case the firm is a local monopoly, or it may involve a national market if it covers all residents of a country. A postal service is another example of a monopoly with a national market. On the other hand, Microsoft Corporation has a monopoly in operating system software with Windows, with roughly 75% of laptops and desktops running Windows in an international market. In all these cases, there is a single firm providing the entire, or the largest part of the market; and it is extremely difficult for other firms to enter the industry. Being the sole producer, the monopolist has significant control over the price of its good or service. Monopolies are not frequently encountered in the real world.

Characteristics of monopolistic competition

- There is a fairly large number of small (or medium-sized) firms in the industry.
- There are no barriers to entry; any new firm can enter the industry.
- There is product differentiation; each firm tries to make its product different from those of the other firms in the industry in terms of various characteristics like the quality, servicing or packaging.

Examples of monopolistic competition include the shoe, clothing, detergent, computer, publishing, furniture and restaurant industries. Monopolistic competition is like perfect competition in that there are many firms in the industry. It differs from perfect competition mainly because of product differentiation. By trying to make its good or service different from any other, each firm tries to be like a little monopoly. This is possible because each firm is the only producer of that particular version of the product. For example, Nike is a monopolist of Nike shoes, and Reebok is a monopolist of Reebok shoes. These firms use product differentiation to gain some control over the price of their products. However, the existence of other similar products (such as other brands of shoes) limits the degree of its power to control the market prices.

Characteristics of oligopoly

- There is a small number of large firms in the industry; because of their small number, the firms are *interdependent*, because the actions of one firm affect the others. This means that each firm tries to predict what the rival firms will do.
- Products may be either differentiated (cars) or undifferentiated (oil).
- There are high barriers to entry; it is difficult for a new firm to enter the industry.

Examples of oligopolies include the car industry, airlines, electrical appliances, electronic equipment (differentiated products) and the oil, steel, aluminum, copper and cement industries (identical or undifferentiated products). Oligopolies have significant control over price. Most industries in the real world are monopolistically competitive or oligopolistic.

Market power and competition

As you learned in Chapter 2, Section *The meaning of competitive markets,* competition occurs when there are many buyers and sellers acting independently, so that no one has the ability to influence the price of a product. Therefore, by definition, market power is the opposite of competition. This can be seen in Table 7.1 which summarises the characteristics of the four market structures, listed in order of increasing market power. The greater the market power, the lower the competition, and the greater the control over price.

Why we study perfect competition

We have seen that perfect competition is a highly unrealistic market structure. Why then do we bother to study it?

It is because it is the only market structure where allocative efficiency is achieved, in other words the only one where the market does not fail. Also, perfect competition sheds light on the benefits of competition in a market economy. Even though competition may not be perfect in the real world, as long as it exists in some form it offers important benefits to the workings of the market system. Perfect competition therefore provides a standard that we can use to evaluate the other three market structures.

Yet firms in perfect competition are also subject to important limitations. By contrast, the other three market structures, while not achieving allocative efficiency, often offer advantages which perfect competition cannot provide. It is therefore important to examine the advantages and disadvantages of each of the four market structures.

TEST YOUR UNDERSTANDING 7.1

1. Explain the relationship between competition and market power, and how the four market structures relate to each other with respect to these features.
2. Define each of the market structures on the basis of their key characteristics.
3. Identify examples of industries that belong to each of the four market structures.

	Number of firms	Type of product	Ease of entry into the industry	Examples	Market power/ control over price	Degree of competition
Perfect competition	very many, very small	homogeneous/ undifferentiated	very easy	agriculture	none	perfect
Monopolistic competition	relatively many, relatively small	differentiated	very easy	restaurants, computer games, books, furniture	some	a good amount
Oligopoly	few large	differentiated or undifferentiated	very difficult	steel, aluminum, cars, household appliances	significant	some (but far less than monopolistic competition)
Monopoly	one, large	one product with no close substitutes	difficult to impossible	public utilities	very significant	none

Table 7.1: Characteristics of market structures listed in order of increasing market power

7.2 Profit maximisation by the rational producer

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain cost, revenue and profit concepts (AO2)
  - total revenue, average revenue, marginal revenue
  - total cost, average cost, marginal cost
  - abnormal profit, normal profit, loss
- calculations from data of profit, marginal cost, marginal revenue, average cost, average revenue (AO4)

To understand market structures, we must first examine some important concepts relating to revenues, costs and profit.

Revenues

Total revenue, average revenue and marginal revenue

Revenues are the payments firms receive when they sell the goods and services they produce. There are three fundamental revenue concepts: total, average and marginal revenue.

The firm's **total revenue** (TR) is obtained by multiplying the price at which a good is sold (P) by the number of units of the good sold (Q):

$$TR = P \times Q$$

The firm's **average revenue** (AR) is revenue per unit of output sold, or total revenue (TR) divided by units of output (Q):

$$AR = \frac{TR}{Q}$$

Note that AR *is always equal to* P, or the price of the product. The reason is that since

$$TR = P \times Q,\ P = \frac{TR}{Q}, \text{ therefore } P = AR.$$

The firm's **marginal revenue** (MR) is the additional revenue arising from the sale of an additional unit of output.

$$MR = \frac{\Delta TR}{\Delta Q}$$

Revenues in two different types of markets

Whereas the *definitions* of revenues apply to all firms, the *analysis* of revenues is not the same, because this depends on whether or not the firm has any ability to control its price. We must therefore make a distinction regarding revenues where:

- the firm is unable to control price; price is constant as output varies: perfect competition
- the firm has control over price; price varies with output: monopolistic competition, oligopoly, monopoly.

Revenue curves for the firm that cannot control price: perfect competition

Table 7.2 shows total, marginal and average revenue based on information on the price and quantity of the good in situations where a firm cannot control price. Columns 3, 4 and 5 are calculated from the data in columns 1 and 2, using the definitions given above. Note that *the price at which the good is sold does not change*; this occurs only under perfect competition. Figure 7.1 plots the data of Table 7.2.

1 Units of output (Q)	2 Product price (P) (€)	3 Total revenue $TR = P \times Q$ (€)	4 Marginal revenue $MR = \frac{\Delta TR}{\Delta Q}$ (€)	5 Average revenue $MR = \frac{TR}{Q}$ (€)
0	10	–	–	–
1	10	10	10	10
2	10	20	10	10
3	10	30	10	10
4	10	40	10	10
5	10	50	10	10
6	10	60	10	10
7	10	70	10	10

Table 7.2: Calculating total, marginal and average revenue when price is constant: the firm has no control price; the case of perfect competition

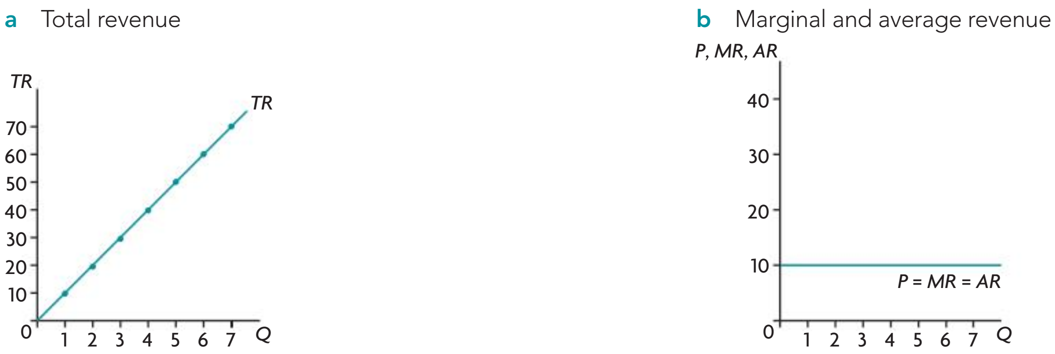


Figure 7.1: Total, marginal and average revenue curves when price is constant: the firm has no control over price; the case of perfect competition

Revenue curves for the firm that can control price: monopoly, monopolistic competition, oligopoly

Table 7.3 shows total, marginal and average revenue from price and quantity information in the case where the firm has some influence over price. The method of calculation is exactly the same as in the competitive case above, where columns 3–5 are calculated using the information of the first two columns. The difference is in the price data, appearing in column 2, showing that *the price at which the good is sold changes as the quantity of output changes*. The lower the price the greater the quantity of output, illustrating the law of demand. This occurs under all market models that we will study other than perfect competition. Figure 7.2 plots the data of Table 7.3.

1 Units of output (Q)	2 Product price (P) (€)	3 Total revenue ($TR = P \times Q$) (€)	4 Marginal revenue $MR = \frac{\Delta TR}{\Delta Q}$ (€)	5 Average revenue $AR = \frac{TR}{Q}$ (€)
0	–	–	–	–
1	12	12	12	12
2	11	22	10	11
3	10	30	8	10
4	9	36	6	9
5	8	40	4	8
6	7	42	2	7
7	6	42	0	6
8	5	40	–2	5
9	4	36	–4	4
10	3	30	–6	3

Table 7.3: Calculating total, marginal and average revenue when price varies: the firm has some control over price; the cases of monopoly, monopolistic competition, oligopoly

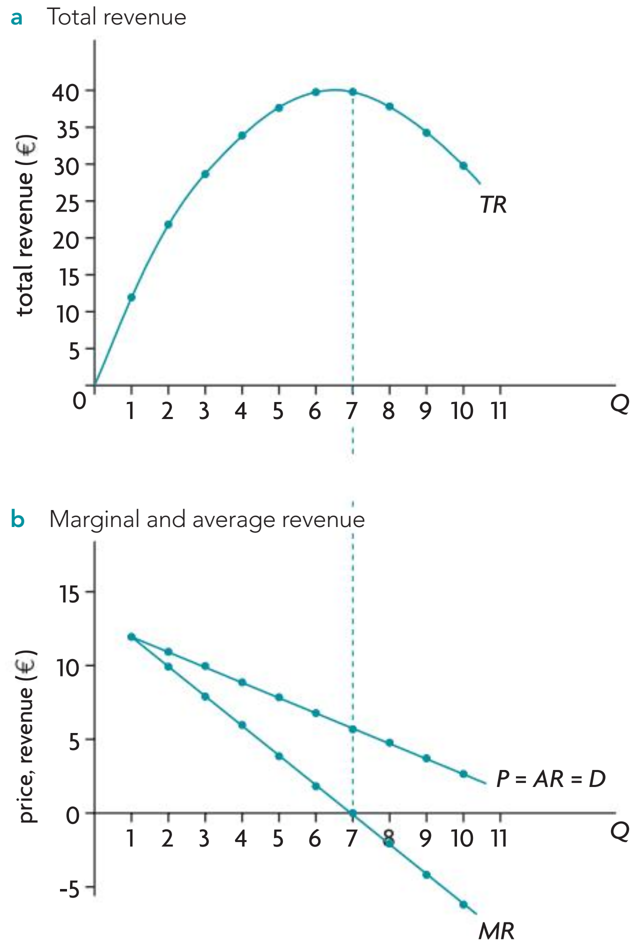


Figure 7.2: Total, marginal and average revenue curves when price varies: the firm has some control over price; the cases of monopoly, monopolistic competition, oligopoly

Calculating revenues

To calculate revenues, it is only necessary to remember and understand the relationships between the various revenue concepts, beginning with the idea that $TR = P \times Q$. If you are given data on P and Q, you can find TR, and from there you can calculate $AR = \frac{TR}{Q}$ and $MR = \frac{\Delta TR}{\Delta Q}$. You can also work backwards to find TR and MR if you know AR and Q, or to find TR and AR if you know MR and Q. Remember also that $AR = P$ in all cases, so that if you know AR, you also know the price of the product.

To calculate revenues from a diagram, we simply read off the information appearing in the graph, and apply exactly the same principles as above to calculate the revenue variable or variables we are interested in.

Costs of production

The meaning of economic costs

When firms use inputs or resources to produce, they incur **costs of production**, which include money payments to buy resources plus anything else given up by a firm for the use of resources. The resources include land, labour, capital and entrepreneurship (see Chapter 1).

When the firm uses resources it does not own, it buys them from outsiders and makes payments of money to the resource suppliers. For example, a firm hires labour and pays a wage; it purchases materials and pays the price to the seller. Such payments made by a firm to outsiders to acquire resources for use in production are known as *explicit costs*.

On the other hand, when the firm uses resources it owns, there is still a cost, which consists of the income that is sacrificed when the firm uses such a self-owned resource. For example, in the case of an office building owned and used by the firm, the cost is the rental income that could have been earned if the building were rented out. You can think of this as an opportunity cost. The hours of work a firm owner puts into his or her own business have an opportunity cost equal to what the firm owner could have earned if s/he had worked elsewhere. The entrepreneurial abilities the firm owner puts into the business (risk-taking, innovative, organisational and managerial abilities) involve a further opportunity cost equal to what these abilities could have earned elsewhere. The sacrificed income arising from the use of self-owned resources by a firm is an *implicit cost*.

TEST YOUR UNDERSTANDING 7.2

1 Define

a total revenue,

b marginal revenue, and

c average revenue.

2 Given the following price and quantity data for a product, calculate total revenue, marginal revenue and average revenue.

Price ($)	5	5	5	5	5
Quantity (thousand units)	0	1	2	3	4

3 Given the following price and quantity data for a product, calculate total revenue, marginal revenue and average revenue.

Price ($)	8	7	6	5	4	3	2
Quantity (thousand units)	2	3	4	5	6	7	8

4 a Outline what can be concluded about how price changes (or does not change) for each unit of output sold in questions 2 and 3.

b Explain the relationship between price and average revenue.

5 Given the following data, calculate total revenue and marginal revenue for each level of output. Identify the price at each level of output.

Quantity (thousand units)	1	2	3	4	5	6
Average revenue (€)	20	18	16	14	12	10

6 Given the following data, calculate total revenue and average revenue for each level of output. Identify the price at each level of output.

Quantity (thousand units)	1	2	3	4	5	6
Marginal revenue (£)	14	12	10	8	6	4

The sum of explicit and implicit costs incurred by a firm for its use of resources, whether purchased or self-owned, are known as *economic costs*. When economists refer to 'costs' they mean 'economic costs'.

Costs of production in the short run

In Chapter 2 (Section *Assumptions underlying the law of supply*), you were introduced to costs of production. (It is suggested that you reread this section before continuing further.) Very briefly, the relevant parts from that section include the following concepts:

- short run = the period of time when at least one factor of production is fixed
- long run = the period of time when all factors of production are variable
- total cost = all costs of production incurred by a firm
- marginal cost (MC) = the extra or additional cost of producing one more unit of output.

To calculate MC, we use the formula

$$MC = \frac{\Delta TC}{\Delta Q}.$$

We now need to introduce one more cost concept: **average cost** (AC), which is cost per unit of output produced, or total cost (TC) divided by units of output (Q):

$$AC = \frac{TC}{Q}$$

Table 7.4 below is the same as Table 2.5 in Chapter 2 (where you learned how to calculate marginal cost) and in addition includes a new column showing *average cost*.

The relationship between average cost and marginal cost in the short run

Figure 7.3 plots marginal cost and average cost based on the data that appear in Table 7.4. As explained in Chapter 2 the MC curve is based on the law of diminishing marginal returns (Section *Assumptions underlying the law of supply*). The same applies to the AC curve. You may recall that diminishing returns holds only in the short run when at least one factor of production is fixed. Therefore *our analysis of costs in this section clearly applies to the short run.*

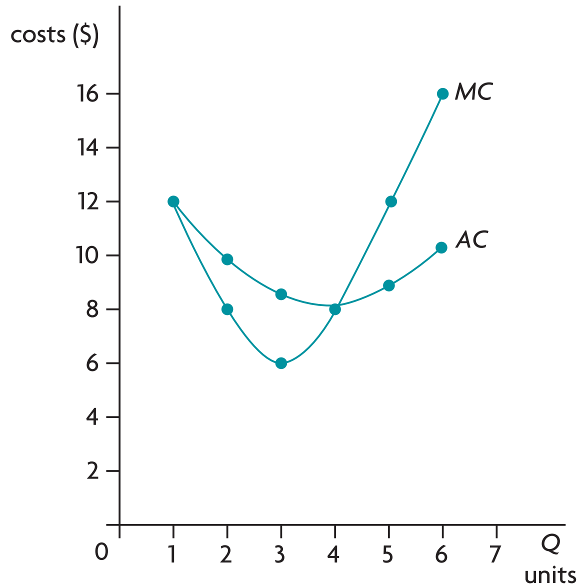


Figure 7.3: Marginal cost and average cost in the short run

Total product (number of units, Q)	Total cost (TC) ($)	Marginal cost (MC) ($)	Average cost (AC) ($)
1	12	12	12
2	20	8	10
3	26	6	8.67
4	34	8	8.5
5	46	12	9.2
6	62	16	10.33

Table 7.4: Total cost, marginal cost and average cost

It is important to note the relationship between the average cost and marginal cost curves: when the marginal cost curve lies below the average cost curve, ($MC<AC$), average cost is falling; and when the marginal cost curve lies above the average costs curve ($MC>AC$), average cost is increasing. *This means that the marginal cost curve always intersects the average cost curve when this is at its minimum*. The reason lies in the mathematical relationship between the average and marginal values of any variable.

Consider a simple example involving test scores. Say you have an average of 80 in your tests. If your next test score (the 'marginal' score) is greater than your average of 80, your average will increase. If your next test score is lower than your average of 80, your average will fall. The relationship between average and marginal test scores is exactly the same as the relationship between average and marginal costs.

Calculating short-run costs of production

You can calculate costs easily if you remember and understand the relationships between the cost concepts we have just discussed. If you have information on units of output (Q) and total cost (TC), you can find $AC = \frac{TC}{Q}$ and $MC = \frac{\Delta TC}{\Delta Q}$. You can also work backwards to find TC and MC if you know AC and Q, or to find TC and AC if you know MC and Q.

To calculate costs from a diagram, you can just read off the information appearing in the graph, and apply exactly the same principles as above to calculate the cost variable or variables you are interested in.

TEST YOUR UNDERSTANDING 7.3

1 Define

a total cost,

b marginal cost, and

c average cost.

2 Given the following quantity and total cost data for a product, calculate average cost and marginal cost.

Quantity (thousand units)	2	5	9	14	18	21	23	24
Total cost (€)	300	400	500	600	700	800	900	1000

CONTINUED

3 Given the following data, calculate total cost and marginal cost for each level of output.

Quantity (thousand units)	1	2	3	4	5
Average cost (€)	180	140	133	135	140

4 Given the following data, calculate total cost and average cost for each level of output.

Quantity (thousand units)	1	2	3	4	5
Marginal cost (£)	14	12	10	12	14

Costs of production in the long run and economies of scale

You may remember that the long run is the period of time when the firm varies (changes) all its factors of production. We are interested in examining how the firm's average costs, or its costs per unit of output, change when it grows larger by *increasing all of its factors of production*.

At any moment in time, when the firm is in the short run, it has a particular short-run AC curve; we can call this its $SRAC$. When it grows over time, we can think of it as going into the long run, increasing all its factors of production, and then going into a new short run with a new $SRAC$. Imagine this process repeating itself again and again: the firm goes from one short-run position to another and then to another, and these short run positions are connected to each other through moments when the firm enters the long run in order to increase all of its factors of production. This process is shown in Figure 7.4(a), where we see that as the firm grows bigger, it has a series of $SRAC$ curves, $SRAC_1$, $SRAC_2$, and so on. Imagine now an infinite number of $SRAC$s; they will trace out the *long run average cost curve*, or $LRAC$, which is the curve that just touches (is tangent to) each of the short-run curves.

We can see that the $SRAC$ curves keeps shifting to the right, which is what we would expect since there is more and more output (Q) being produced. But note that the $SRAC$ curves not only move to the right, but also at first move downward, and after a point they move upward. It is this U-shape of the $LRAC$ curve that we want to examine.

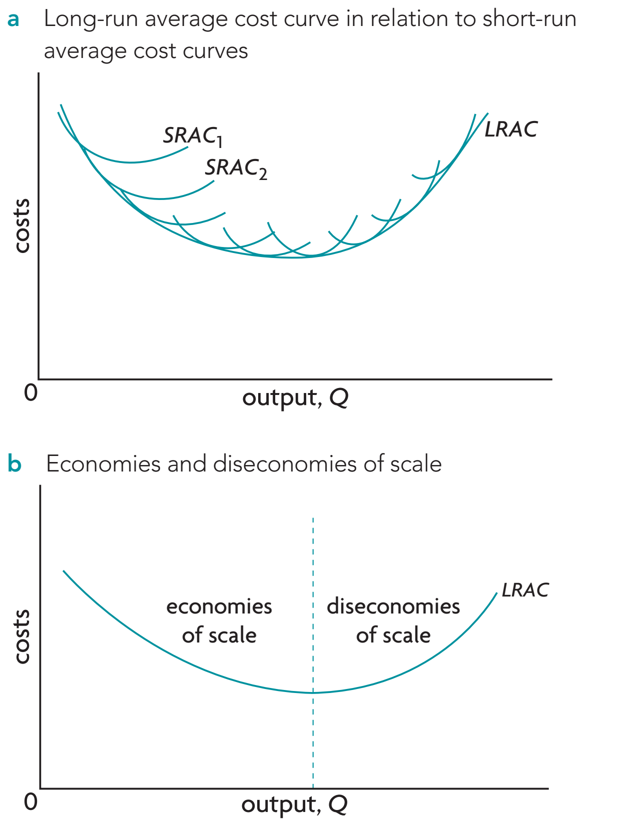


Figure 7.4: The long-run average cost curve

The reasons for the U-shape of the *LRAC* have nothing to do with diminishing marginal returns, which are a feature of short-run production and costs. The U-shape of the *LRAC* curve can be found in *economics and diseconomies of scale*.

Economies of scale are decreases in average costs of production over the long run as a firm increases all its factors of production. They explain the downward-sloping portion of the *LRAC* curve.

Some reasons why this occurs include:

- **Specialisation of labour.** As the scale of production increases, more workers must be employed, allowing for greater labour *specialisation*. Each worker specialises in performing tasks that make use of skills, interests and talents, thus increasing efficiency and allowing output to be produced at a lower average cost.
- **Specialisation of management.** Larger scales of production allow for more managers to be employed, each of whom can be specialised in a particular area (such as production, sales, finance, and so on), again resulting in greater efficiency and lower average cost.
- **Bulk buying of inputs (factors of production)**. As quantities of inputs purchased increase, the price per unit drops.
- **Financing economies**. Larger firms may have lower interest rates, thus contributing to lower costs per unit of output.
- **Spreading of certain costs, such as marketing, over larger volumes of output.** Costs of certain activities such as marketing and advertising, design, research and development, result in lower average costs if they can be spread over large volumes of output.

Diseconomies of scale, on the other hand, are increases in the average costs of production in the long run as a firm increases its output by increasing all its inputs. They are responsible for the upward-sloping part of the *LRAC* curve: as a firm increases its scale of production, average costs increase.

Reasons for diseconomies of scale include:

- **Co-ordination and monitoring difficulties.** As a firm grows larger, its management may run into difficulties of co-ordination, organisation, co-operation and monitoring. The result involves growing inefficiencies causing average costs to increase as the firm expands.
- **Communication difficulties.** A larger firm size may lead to difficulties in communication between various component parts of the firm, resulting in inefficiencies and higher average costs.
- **Poor worker motivation.** If workers begin to lose their motivation, to feel bored and to care little about their work, they become less efficient, resulting in higher average costs.

Profits

The meaning of profit

Standard economic theory assumes that firms display rational behaviour by trying to maximise their profits, meaning that they try to make their profits as large as possible (Chapter 2). In general,

profit = revenues – costs of production

In economics, profit is defined as:

profit = total revenue – economic costs
= total revenue – the sum of explicit costs + implicit costs

since economic costs are the sum of explicit plus implicit costs.

Profit maximisation

Standard economic theory of the firm assumes that firm behaviour is guided by the firm's goal to maximise profit. **Profit maximisation** involves determining the level of output that the firm should produce to make profit as large as possible.

Yet firms do not always make a profit; in some cases, their total revenue is not sufficient to cover all costs, in which case they make a **loss**, which can be thought of as negative profit. If a firm is making a loss, it may eventually go out of business, but until it decides to shut down, it will be interested in producing the quantity of output that will make its loss as small as possible. Therefore, the theory of the firm is also concerned with how much output a loss-making firm should produce in order to minimise its loss.

There are two approaches to analysing profit maximisation (or loss minimisation): one involves use of *total revenues and total costs* and the other involves use of *marginal revenues and marginal costs*. Both these approaches yield the same results for the profit-maximising (or loss-minimising) level of output.

Profit maximisation based on the total revenue and total cost approach

This approach is based on the simple principle that

profit = total revenue (TR) − total cost (TC)

where TC is the firm's *economic costs* (explicit plus implicit).

The firm's profit-maximisation rule is to produce the level of output where $TR - TC$ is as large as possible.

The amount of profit made by the firm is equal to the numerical difference between TR and TC.

> Positive profit: $TR > TC$; the firm earns abnormal profit.
>
> Zero profit: $TR = TC$; the firm earns normal profit.
>
> Negative profit: $TR < TC$; the firm makes a loss

We will return to the terms *abnormal profit* and *normal profit*.

Profit maximisation based on the marginal revenue and marginal cost approach

Profit maximisation using this approach is based on a comparison of marginal revenue (MR) with marginal cost (MC) to determine the profit-maximising level of output. The firm's profit-maximisation rule (and loss-minimisation) is to choose to produce the level of output where $MC = MR$.

In Figure 7.5, both parts (a) and (b) show the standard MC curve that we studied above (as well as in Chapter 2 in connection with the firm's supply curve). As we know, there are two kinds of marginal revenue curves, depending on whether or not the firm has control over the price of its output. Part (a) shows the MR curve of the firm with no control over price. Part (b) shows the MR curve of the firm with some control over price. Both parts (a) and (b) illustrate the identical principle about profit maximisation.

According to the profit-maximising rule, $MC = MR$, the point of intersection between the MC and MR curves determines the profit-maximising level of output; this is $Q_{\pi max}$ in Figure 7.5(a) and (b). Why is this so? Consider a firm producing output Q_1 in both parts (a) and (b), where $MR > MC$. If this firm increases its output by one unit, the additional revenue it would receive (MR) will be greater

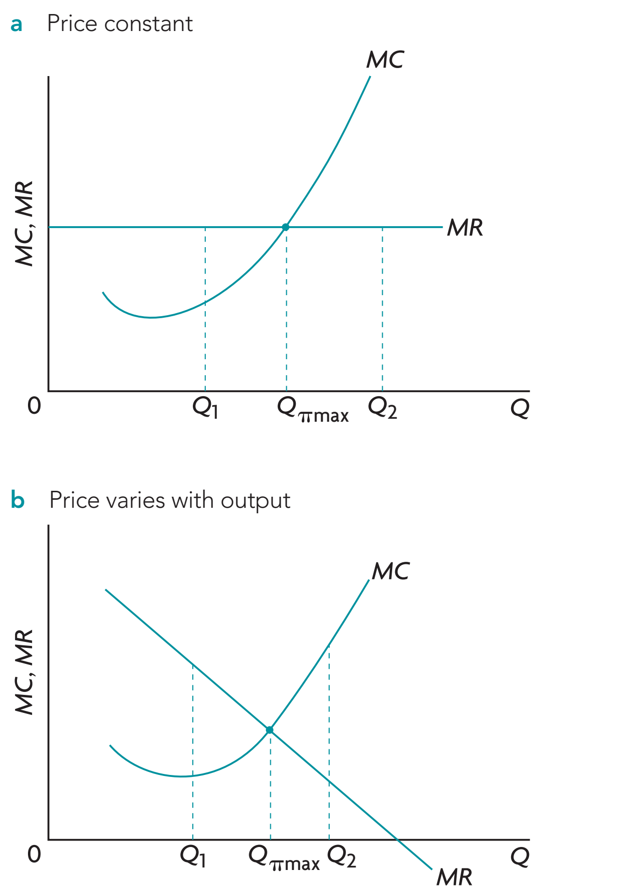


Figure 7.5: Profit maximisation using the marginal revenue and marginal cost approach

than its additional cost (MC). It is therefore in the firm's interests to increase its level of output until it reaches $Q_{\pi max}$ where $MR = MC$. If it continues to increase output beyond $Q_{\pi max}$, say to Q_2, where $MR < MC$, the additional revenue it would receive for an extra unit of output is less than the additional cost, and so it should cut back on its Q. There is only one point where the firm can do nothing to improve its position, and that is $Q_{\pi max}$, where $MR = MC$, and profit is the greatest it can be.

Firms maximise profits or minimise losses when they produce a quantity of output where $MC = MR$. This holds for all market structures.

When we are given data for MC and MR (and no information on total costs and total revenues), all we can do is find the profit-maximising level of output where $MC = MR$, but we cannot find the amount of profit (or loss) unless we have more information. We will see how this is done below when we study the market structures.

TEST YOUR UNDERSTANDING 7.4

1 a Identify the two approaches to profit maximisation by firms.

b Outline the profit-maximising rule of firms in each of the two approaches.

2 a Say a firm is producing a level of output Q where $MC > MR$. State what it should do to increase its profit (or reduce its loss).

b State what it should do if it is producing Q where $MC < MR$.

3 Assume that a firm that has no control over its price sells its output at $5 per unit.

a Given the following data, use the total revenue and total cost approach to determine the level of output at which the firm will maximise profit.

b Calculate how much profit it will make.*

c Calculate the amount of profit (or loss) when $Q = 3$, $Q = 6$, $Q = 10$.

Units of output (Q)	1	2	3	4	5	6	7	8	9	10
Total cost ($)	15	18	20	21	23	26	30	35	41	48

CONTINUED

4 Given the data in question 3

a determine the level of output at which the firm will maximise profit using the marginal revenue (MR) and marginal cost (MC) approach. (*Hint:* you must use the information in the question to find MR and MC.)

b Did you find the same profit-maximising level of output as in question 3(a)?

5 Suppose that a firm with some control over price faces the costs and prices per unit of output shown in the table below.

a Use the total revenue and total cost approach to determine the level of output at which the firm will maximise profit.

b Calculate how much profit will it make.*

c Calculate the amount of profit (or loss) when $Q = 2$, $Q = 3$, $Q = 8$.

Units of output (Q)	1	2	3	4	5	6	7	8
Total cost ($)	15	18	20	21	23	26	30	35
Price ($)	10	9	8	7	6	5	4	3

6 Given the data in question 5,

a determine the level of output at which the firm will maximise profit using the marginal revenue (MR) and marginal cost (MC) approach. (*Hint:* you must use the information in the question to find MR and MC.)

b Did you find the same profit-maximising level of output as in question 5?

*When using the *TR* and *TC* approach your results give two profit-maximising levels of output, whereas the *MR* and *MC* approach gives only one. This is because the *MR* and *MC* approach is actually more precise than the *TR* and *TC* approach. It is a good idea to use the larger of the two values of output that you get by using the *TR* and *TC* approach.

The meaning of normal profit

When profit is equal to zero, and total revenue is equal to total economic costs, the firm is said to be making normal profit.

> *Normal profit* can be defined as the minimum amount of revenue that the firm must receive so that it will keep the business running (as opposed to shutting down). It can also be defined as the amount of revenue that covers all explicit and implicit costs. Therefore, a firm earns normal profit when total revenue = economic costs and profit = zero.

These apparently different definitions are in fact consistent: the minimum amount of revenue the firm must receive to make it worthwhile to stay in business is equal to the revenue that covers all of the firm's costs, implicit plus explicit.

Note that normal profit also *includes the payment for entrepreneurship*. Entrepreneurship, you will remember, includes the talents to organise and manage a business and take risks. Entrepreneurship receives a payment just as all other factors of production do, and this payment is included in normal profit. This means that if you are the owner of a pizza restaurant which is earning normal profit (meaning zero profit) you are still getting paid for all your work as entrepreneur in your business. Therefore you have no reason to close down your restaurant.

This means that a firm will continue to produce even if it is earning zero profit, meaning it is earning normal profit.

TEST YOUR UNDERSTANDING 7.5

1. Define profit and normal profit, and explain the difference between them.
2. Explain why profit can be positive, zero or negative.
3. A firm earns zero profit, and yet it does not shut down. Explain why.

7.3 Perfect competition

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain firms in perfect competition as price takers having no market power (AO2)
- explain profit maximisation in the short run and in the long run (AO2)
- explain the meaning of allocative efficiency in terms of its necessary conditions, $P = MC$ or $MB = MC$ or maximum social surplus (AO2)
- draw diagrams showing (AO4)
  - the perfectly competitive firm as a price taker where $P = D = AR = MR$
  - the perfectly competitive firm making abnormal profit, normal profit, loss
  - perfectly competitive market equilibrium showing allocative efficiency
- discuss the advantages and disadvantages of perfect competition (AO3)

Let's remind ourselves of the characteristics of perfect competition:

- There is a very large number of firms in the industry.
- All the firms in the industry sell a homogeneous or identical (undifferentiated) product.
- There are no barriers to entry into the industry.

Demand and revenue curves

The demand curve (average revenue curve) facing the firm

Suppose you have a farm that produces strawberries. You are one of many small strawberry producers, your strawberries are very similar to those produced by other strawberry farmers, and anyone who would like to produce strawberries can do so (there are no barriers to entry).

Table 7.5 summarises the cost, product, revenue and profit concepts we have studied.

Revenue concepts	Definition	Formula
Total revenue	The total earnings of a firm from the sale of its output.	$TR = P \times Q$
Marginal revenue	The additional revenue of a firm arising from the sale of an additional unit of output.	$MR = \frac{\Delta TR}{\Delta Q}$
Average revenue	Revenue per unit of output.	$AR = \frac{TR}{Q} = P$
Cost concepts		
Explicit cost	The monetary payment made by a firm to an outsider to acquire an input.	
Implicit cost	The income sacrificed by a firm that uses a resource it owns.	
Total cost (*TC*)	The sum of explicit and implicit costs	
Average total cost (*AC*)	Total cost per unit of output.	$AC = \frac{TC}{Q}$
Marginal cost (*MC*)	The change in cost arising from one additional unit of output.	$MC = \frac{\Delta TC}{\Delta Q}$
Long-run average cost (*LRAC*) curve	A U-shaped curve showing average costs in the long run when all of the firm's inputs are variable.	
Profit concepts		
Profit	Total revenue minus total cost (the sum of explicit plus implicit costs).	TR−TC
Normal profit	Occurs when total revenue equals total cost. It is the minimum amount of revenue required by a firm to keep running.	TR=TC
Abnormal profit	Profit that results when total revenue is greater than total cost. It is revenue that is over and above normal profit.	TR>TC
Loss	Negative profit; occurs when total revenue is less than total cost	TR<TC

Table 7.5: Summary of cost, product, revenue and profit concepts

Figure 7.6(a) shows standard market demand and supply curves for strawberries, which determine the equilibrium price, P_e. Figure 7.6(b) shows the demand curve for strawberries *as it appears to you, the strawberry producer*. It is perfectly elastic, appearing as a horizontal line at P_e determined in the market. A perfectly elastic demand curve has a price elasticity of demand (*PED*) equal to infinity throughout its range (see Chapter 3, Section 3.1). What does this mean for you?

As a strawberry producer, being small, you can do nothing to influence this price; you must accept P_e and sell the amount of strawberry output that will

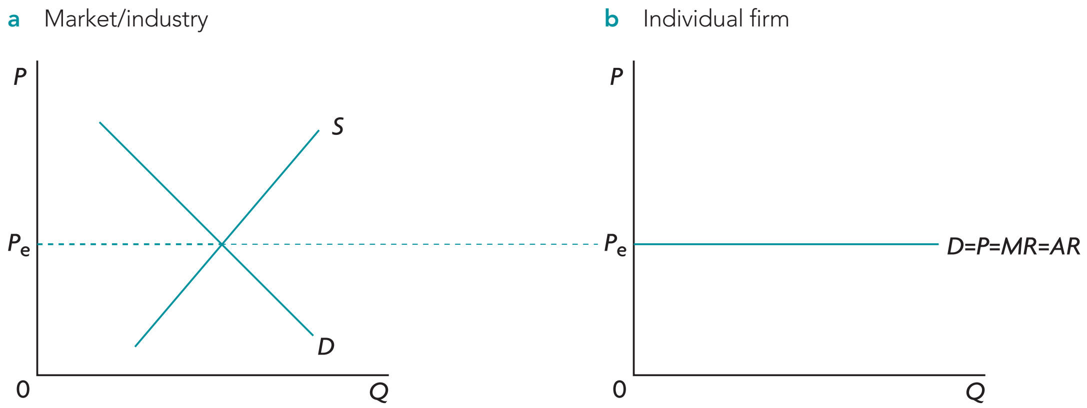


Figure 7.6: Market (industry) demand and supply determine demand faced by the perfectly competitive firm

maximise your profit. Your farm (or your firm) is therefore a **price-taker**. If you raise your price above P_e, you will not sell any strawberries because buyers will buy strawberries elsewhere at the lower price P_e. On the other hand, since you can sell all you want at price P_e, you would have nothing to gain and something to lose (some revenue) if you dropped your price below P_e. Therefore, you sell all your strawberry output at P_e.

> The demand curve for a good facing the perfectly competitive firm is perfectly elastic (horizontal) at the price determined in the market for that good. This means the firm is a *price-taker*, as it accepts the price determined in the market. The firm has no ability to influence price therefore it has no market power.

The firm's revenue curves

The firm we are considering is the one studied in Section 7.2 above, when we studied revenue data and curves for the firm that is *unable to influence price*. Consider once again the example used in Table 7.2 and Figure 7.1. Assume that a perfectly competitive firm sells a good at €10 per unit. In Table 7.2, column 3 shows total revenue, calculated by multiplying units of output in column 1 by price shown in column 2. Column 4 calculates marginal revenue, by taking the change in total revenue and dividing it by the change in output. Column 5 shows average revenue, obtained by dividing total revenue by quantity of output. The data in the table reveal an interesting pattern:

> No matter how much output the perfectly competitive firm sells, $P = MR = AR$ and these are constant at the level of the horizontal demand curve. This follows from the fact that price is constant regardless of the level of output sold.

This result holds only for firms operating under perfect competition, because these are the only firms that are unable to influence price and are forced to sell all their output at the single price determined in the market.

The data of Table 7.2 are plotted in Figure 7.7 below which is the same as Figure 7.1(b), where we see that since price is constant at €10, $P = MR = AR$, and they all coincide with the horizontal demand curve.

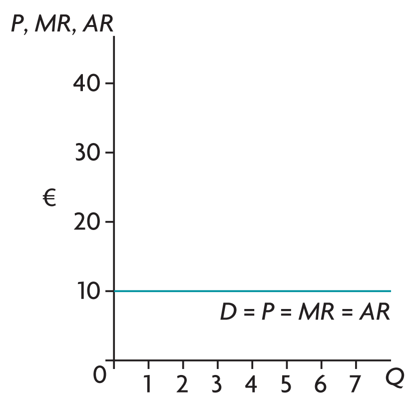


Figure 7.7: Demand, marginal revenue and average revenue in perfect competition

TEST YOUR UNDERSTANDING 7.6

1. Outline the characteristics of perfect competition.
2. a. Explain why the perfectly competitive firm is a price-taker.
   b. Outline what would happen if this firm tried to raise its price above the market price or if it lowered its price below the market price.
3. Outline how the demand curve facing the perfectly competitive firm relates to the industry/market equilibrium.
4. a. Using a diagram, explain the relationship between the firm's average revenue (*AR*) and marginal revenue (*MR*) in perfect competition. Outline how are they related to
   b. product price,
   c. the demand curve facing the firm, and
   d. the principle that each firm is a price-taker.

Profit maximisation in the short run

Remember, the short run is the period when the firm has at least one fixed input. This means the number of firms in the industry is also fixed. To enter or leave an industry, a firm must be able to vary *all its inputs*. Since this cannot be done in the short run, firms cannot enter or leave the industry (until they move into the long run).

In terms of your strawberry farm, this means that you have some fixed inputs such as the land you cultivate, and perhaps your farm machinery. As long as all other strawberry producers also have a fixed amount of agricultural land and farm machinery, no one can enter or leave strawberry production.

When a firm wants to maximise profit in the short run, what must it do? Since it is a price-taker, it cannot influence its selling price. It can only make a choice on how much quantity of output it should produce. We will see how the firm does this using the marginal revenue and marginal cost rule, introduced above.

Short-run profit maximisation based on the marginal revenue and marginal cost rule

The analysis consists of three steps.

i **Find the point where marginal revenue equals marginal cost to determine profit-maximising (or loss-minimising) level of output.** As we know, a firm interested in maximising profit (or minimising loss) produces output where $MR = MC$. Figure 7.5(a) shows $Q_{\pi\max}$ to be the profit-maximising level of output.

ii **Compare average revenue (or price) and average cost to determine the amount of profit (or loss) per unit of output.** A comparison of average revenue (which is equal to price) with average cost shows the amount of profit (or loss) per unit of output. We know profit = $TR - TC$. If we divide this throughout by output, Q, we get an expression for profit per unit of output, in other words, in terms of averages:

$$\frac{profit}{Q} = \frac{TR}{Q} - \frac{TC}{Q}$$

Alternatively,

$$\frac{profit}{Q} = \text{AR} - \text{AC}.$$

Moreover, since $P = AR$, it follows that

$$\frac{profit}{Q} = \text{P} - \text{AC}$$

This is the key to calculating how much profit or loss per unit of output the firm is making.

iii **Find total profit (or total loss).** To do this, we multiply:

$$\frac{profit}{Q} \text{ by } Q \text{ (or } \frac{loss}{Q} \text{ by } Q\text{)}$$

At the profit-maximising level of output *Q*:

- If *AR* > *AC* (or *P* > *AC*), the firm makes abnormal profit (positive profit).
- If *AR* = *AC* (or *P* = *AC*), the firm makes normal profit (zero profit)
- If *AR* < *AC* (or *P* < *AC*), the firm makes a loss (negative profit).

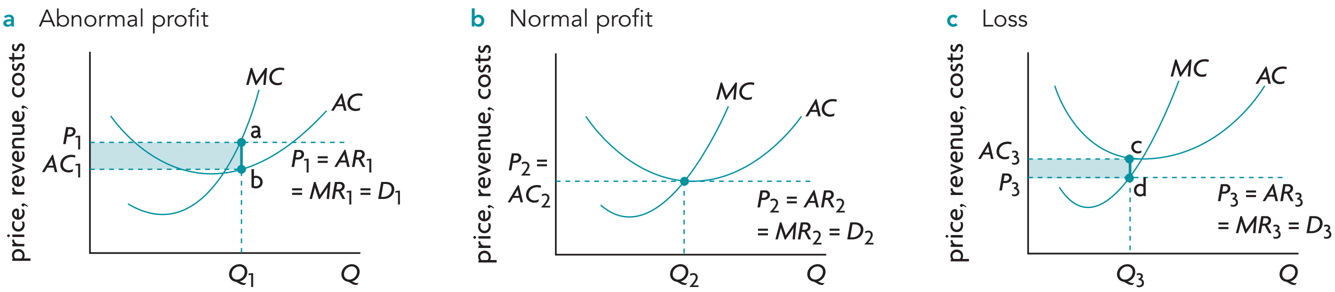


Figure 7.8: Short-run profit maximisation in perfect competition

Using the above three step approach, we will examine the behaviour of the perfectly competitive firm in the short run, making use of the diagrams in Figure 7.8. Each of these diagrams contains identical *AC* and *MC* cost curves; note that *MC* always intersects *AC* at its minimum point. What differs between the diagrams is the position of the perfectly elastic demand curve, showing different possible prices that the firm, being a price-taker, must accept.

Profit maximisation and abnormal profit

In Figure 7.8(a), price $P_1 = AR_1 = MR_1$ represent the demand curve facing the firm. Using the rule $MR = MC$, we arrive at the profit maximising level of output Q_1 (simply draw a vertical line from the point of intersection to the horizontal axis). We then compare P_1 with AC, along this same vertical line, and since $P_1 > AC$, we conclude the firm is making abnormal profit per unit equal to $P_1 - AC_1$, given by the vertical distance between points a and b. To find total profit we multiply profit per unit times the total number of units produced, given by

profit = $\frac{profit}{Q} \times$ Q and is represented by the shaded area in the diagram.

> When *P*>*AC* (or *AR*>*AC*) at the level of output where *MC* = *MR*, the firms earns abnormal profit (positive profit).

Profit maximisation and normal profit

Suppose the market-determined price falls to P_2, corresponding to demand curve D_2 as in Figure 7.8(b). Applying again the $MR = MC$ rule, we find the profit-maximising level of output Q_2. Comparing P_2 with *AC* at output Q_2, we see they are equal to each other; therefore, profit per unit is $P_2 - AC_2 = 0$. Therefore, profit is zero and the firm is earning normal profit. When profit is zero, price equals minimum *AC*. The firm's total revenues are equal to its total costs.

> When *P* = minimum *AC* (or *AR* = minimum *AC*) at the level of output where *MC* = *MR*, the firm earns normal profit (zero profit).

Loss minimisation

If the market price falls below minimum *AC*, such as P_3 in Figure 7.8(c), corresponding to demand curve D_3 as in Figure 7.8(b), the firm does not earn enough revenue to cover all its costs. Using the $MC = MR$ rule, we see that the profit-maximising or loss-minimising level of output is Q_3, at which $P_3 < AC$, indicating the firm is making a negative profit, or loss. Therefore, Q_3 is the firm's loss-minimising output. $AC_3 - P_3$, or the difference between points c and d, represents the firm's loss per unit of output, or $\frac{loss}{Q}$. If we multiply this vertical distance by Q_3, we get the firm's total loss, given by the shaded area.

> When *P* < minimum *AC* (or *AR* < minimum *AC*) at the level of output where *MC* = *MR*, the firm makes a loss (negative profit).

Profit maximisation in the long run

In the long run, all the firm's factors of production are variable; therefore, the number of firms in the industry is no longer unchanging. New firms can enter the industry, existing firms can change their size (increase or decrease all their inputs), or firms can leave the industry altogether. There is therefore free entry of firms in an industry.

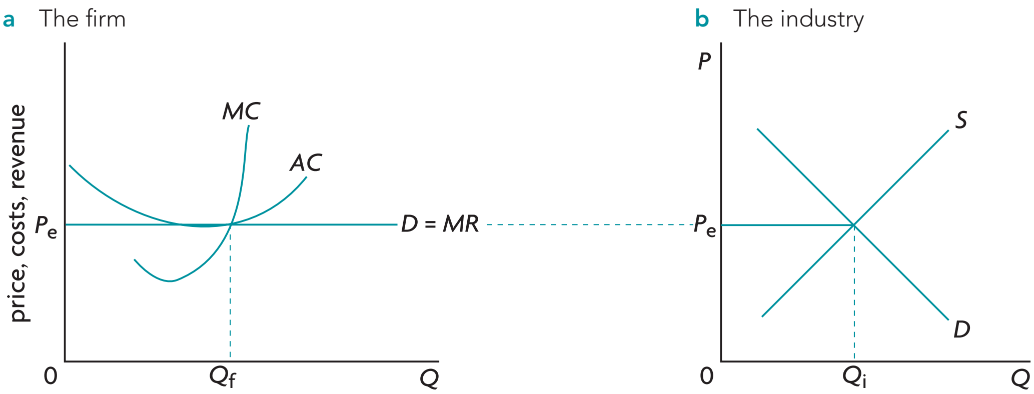


Figure 7.9: The firm and industry long-run equilibrium position in perfect competition

Normal profit in the long run

In the long-run equilibrium of perfect competition, all firms earn zero profit, in other words, they earn normal profit. The long-run equilibrium position of each firm and the industry under perfect competition is shown in Figure 7.9. The market settles at the price P_e, which is just equal to the firm's *minimum AC*, where each firm is earning normal profit, since $P = AC$ (or $AR = AC$). You may note that at this point AC must be minimum since profit maximisation occurs where $MR = MC$, and MC intersects AC at its minimum point. Each firm in the industry produces output Q_f, and the industry as a whole produces output Q_i (equal to the sum of all the firms' outputs).

Why does this happen? How is it that the firms that were in the short run earning abnormal profits, or were making losses, end up making normal profits in the long run?

Suppose the strawberry industry is profitable, meaning that strawberry farms (or firms) are making abnormal profits, so that $P>AC$ (or $AR>AC$). In the long run, when firms can vary all their inputs, new farmers are attracted into the strawberry industry since they would also like to earn abnormal profits. As more and more farms begin to produce strawberries, the supply of strawberries increases, and this has the effect of reducing the price of strawberries. But as the price of strawberries falls, the abnormal profits of strawberry producers fall. The price of strawberries continues to fall until it is just equal to minimum AC. At that point the strawberry producers, who as *price-takers* must accept the price that is determined in the market, end up earning normal profit where $P = AC$ (or $AR = AC$).

Suppose now that the strawberry industry were loss-making, in which case $P<AC$ (or $AR<AC$). In the long run some strawberry farms would close down, or else they might stop producing strawberries and begin producing another more profitable crop. As strawberry farmers leave the industry or switch out of strawberries, the supply of strawberries decreases, causing the price of strawberries to rise. This process continues until price is just equal to AC. Here too, the remaining strawberry producers in the industry will end up earning normal profit where $P = AC$ (or $AR = AC$).

The interested student may see how this process occurs using diagrams (presented as Supplementary material in the digital coursebook).

In perfectly competitive long-run equilibrium, firms' profits and losses are eliminated, and revenues are just enough to cover all costs so that every firm earns normal profit.

This process illustrates an important principle, which is that competition leads to a process of firms opening (or closing) such that price is driven down (or up) to the lowest level that is acceptable to the firm in order for it to continue operating. This is the price that is equal to minimum average cost, and that allows the firm to earn normal profit so that it is just covering all its costs. As we will see in the pages below, *anything that restricts competition results in higher prices, abnormal profits for firms and welfare loss for society*.

TEST YOUR UNDERSTANDING 7.7

1. Using diagrams, show when a firm
   - a earns profit (show profit per unit and total profit),
   - b earns normal profit, and
   - c earns negative profit (a loss) (show loss per unit and total loss).

CONTINUED

2 For each of the following, calculate abnormal profit or loss per unit of output, and then calculate total abnormal profit or loss:

a Q = 200 units; AC = \$8, P = \$9

b Q = 250 units; AC = \$15, P = \$13

c Q = 150 units; AC = \$17, P = \$17

3 Given the information in the table,

a if price = €6, calculate how much the profit-maximising (loss-minimising) firm will produce, and how much profit or loss it will make;

b if price = €4, calculate how much the profit-maximising (loss-minimising) firm will produce, and how much profit or loss it will make.

Total product (units of output)	Average cost (€)	Marginal cost (€)
1	14.00	4
2	8.50	3
3	6.33	2
4	5.00	1
5	4.40	2
6	4.17	3
7	4.14	4
8	4.25	5
9	4.44	6
10	4.70	7

4 a Using a diagram show and explain a perfectly competitive firms' long run equilibrium position.

b Outline the process by which the firm reaches this position.

c Explain what kind of profit the firm makes in long run equilibrium.

Allocative efficiency

In Chapter 2, we saw that competitive markets achieve allocative efficiency, because equilibrium is determined by $MB = MC$, where consumer plus producer (social) surplus is maximum. This discussion focused on efficiency at the level of the *market*, or industry. We now want to see how efficiency can be analysed also at the level of the individual firm.

Review of allocative efficiency

Allocative efficiency is achieved when $MB = MC$; but since $MB = P$, it follows there is allocative efficiency when $P = MC$. Note that this condition holds only when there are no externalities, in which case it is also true that $MSB = MSC$ (this was our condition for allocative efficiency when we studied externalities in Chapters 5 and 6).

> *Allocative efficiency* occurs when firms produce the particular combination of goods and services that consumers mostly prefer. The condition is the following:
>
> Allocative efficiency is achieved when $P = MC$ *(or alternatively $MB = MC$)*

The price, P, paid by consumers to acquire a good reflects the marginal benefit they derive from consumption of one more unit of the good and shows the amount of money they are willing to pay to buy one more unit. Marginal cost, MC, measures the value of the resources used to produce one extra unit of the good. When price is equal to marginal cost, there is equality between what consumers are prepared to pay for one more unit and what it costs to produce it.

What would happen if P and MC were not equal to each other? If $P > MC$, an additional unit of the good is worth more to consumers than the cost to produce it. There is an underallocation of resources to its production, and consumers would be better off if more of it were produced. If $P < MC$, an additional unit of the good costs more to produce than it is worth to consumers; there is an overallocation of resources to the good, and consumers would be better off if output were reduced. In both these cases, allocative inefficiency results. Therefore, resources are allocated efficiently only when the price of a good is equal to the marginal cost of producing it.

Allocative efficiency and perfect competition

Figure 7.10 shows the long-run equilibrium position of a firm and industry in perfect competition. Part (a) shows the firm to be earning normal profit, and indicates that in long-run equilibrium, the perfectly competitive firm achieves allocative efficiency since at the profit-maximising level of output, Q_e, $P = MC$. Part (b) shows the industry to be achieving allocative efficiency ($MB = MC$ and social surplus is maximum). It illustrates how the efficiency at the level of the firm corresponds to efficiency at the level of the industry.

> In long-run equilibrium under perfect competition, the firm achieves allocative efficiency where $P = MC$ (or where $MB = MC$). At the level of the industry, social surplus (consumer plus producer surplus) is maximum, and $MB = MC$.
>
> The achievement of allocative efficiency in long-run equilibrium is an important result, because perfect competition is the only market structure where this occurs.

Evaluating perfect competition

Perfect competition, though not a realistic market structure, offers a number of insights into the workings of competitive markets.

Insights provided by the model

- **Allocative efficiency.** Perfect competition leads to the best or 'optimal' allocation of resources, achieved through $P = MC$ (or $MB = MC$) in long-run equilibrium.
- **Low prices for consumers.** Consumers benefit from low prices, due to the absence of abnormal profits, which would have led to a higher price. You can check this by comparing Figure 7.8(a) and (b), showing that the price when the firm earns abnormal profit is higher than the price when the firm earns only normal profit.
- **Competition leads to the closing down of inefficient producers.** Inefficient firms are those that produce at higher than necessary costs. Inefficiency could be due to factors like less productive labour, or the use of outdated technologies, or poor entrepreneurship. The revenues of inefficient firms are insufficient to cover all costs, leading to losses that force these firms to leave the industry in the long run.
- **The market responds to consumer tastes.** Changes in consumer tastes are reflected in changes in market demand and therefore market price. By creating short-run abnormal profits or losses, price changes result in long-run adjustments that make the quantity of output produced by the industry respond to consumer tastes.

Limitations of the model

- **Unrealistic assumptions.** The model rests on strict and unrealistic assumptions that are rarely met in the real world.
- **Cannot take advantage of economies of scale.** Economies of scale lead to lower average costs as a firm grows larger and larger. In perfect competition firms are too small to grow to a size large enough to have economies of scale.
- **Lack of product variety.** All firms within an industry produce identical or undifferentiated (homogeneous) products, however consumers prefer product variety.

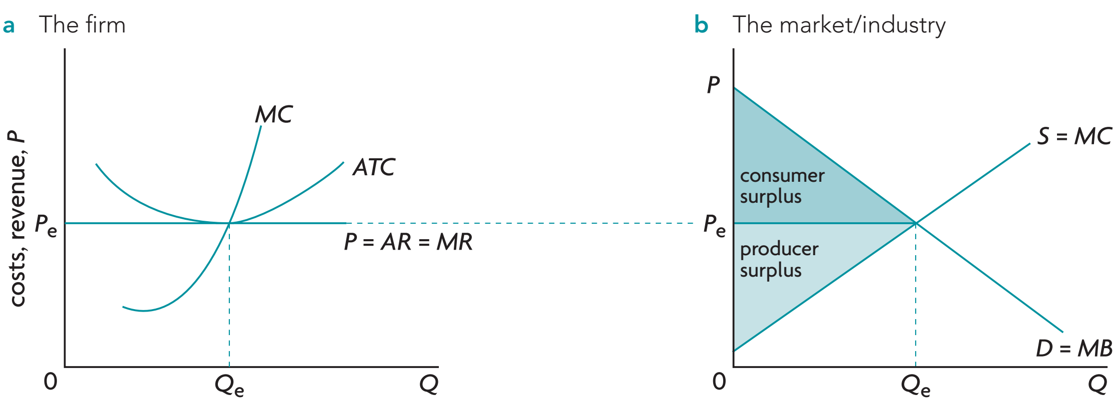


Figure 7.10: Allocative efficiency in perfect competition in the long run

- **Limited ability to engage in new product development.** The lack of abnormal profits in the long run does not offer firms the necessary funds to pursue research and development. In any case, since products are undifferentiated, firms do not have the incentive to pursue product development or improvement.

TEST YOUR UNDERSTANDING 7.8

1 Outline why we study the perfectly competitive market model extensively when this model is based on so many unrealistic assumptions.

2 a Explain the meaning of and state the conditions for allocative efficiency.

b Using diagrams, show how the perfectly competitive *firm* and *industry* achieve allocative efficiency in long-run equilibrium.

3 Evaluate the perfectly competitive market model by referring to the insights it offers and its limitations.

7.4 Monopoly

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the monopolist's market power using a diagram where *AR*>*MC* (AO2, AO4)
- explain profit maximisation by the monopolist with diagrams showing abnormal profit, normal profit, loss (AO2, AO4)
- explain the monopolist's allocative inefficiency illustrating market failure (AO2)
- explain welfare loss compared to perfect competition with a diagram showing lower output and higher price and welfare loss (AO2, AO4)
- explain natural monopoly and draw a diagram for natural monopoly (AO2, AO4)
- discuss advantages and disadvantages of monopoly (AO3)

The term 'monopoly' is derived from the Greek word μονοπώλιο (monopolio) meaning 'single seller'. We begin by reminding ourselves of the characteristics of monopoly:

- there is a single firm in the industry
- the firm produces and sells a unique good or service, with no close substitutes
- there are high barriers to entry in the industry.

Monopoly lies at the opposite extreme of market structures to perfect competition. As a single seller, the monopolist faces no competition from other firms and it has substantial market power (the ability to control price).

Barriers to entry

There are several kinds of barriers to entry, described below.

Economies of scale

Economies of scale result in the downward-sloping portion of a firm's long-run average cost curve (*LRAC*), resulting in lower average costs as the firm increases its size (Figure 7.11). Large economies of scale create a barrier to entry. In Figure 7.11 the average costs of a large firm on $SRAC_1$ are substantially lower than the average costs faced by a smaller firm on $SRAC_2$. The large firm can charge a lower price than the smaller firm, and can force the smaller firm into a situation where it will not be able to cover its costs. Therefore, if new firms try to enter the industry on a small scale they will be unable to compete with the larger one.

On the other hand, a new firm attempting to enter the market on a very large scale would encounter huge start-up costs, and would be unlikely to take the risk. Economies of scale form a significant barrier to entry also in oligopolies.

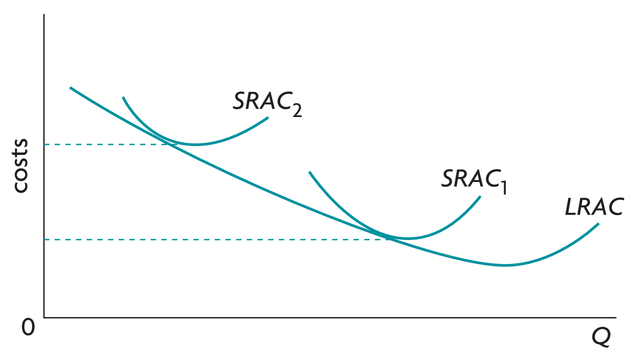


Figure 7.11: Economies of scale as a barrier to entry

Natural monopolies

Natural monopolies are firms that have economies of scale so large that they can produce for an entire market and still not exhaust their economies of scale. They will be examined in more detail below.

Branding

Branding involves the creation by a firm of a unique image and name of a product. It works through advertising campaigns that try to influence consumer tastes in favour of the product, attempting to establish consumer loyalty. If branding is successful, many consumers will be convinced of the product's superiority, and will be unwilling to switch to substitute products, even though these may be qualitatively very similar. Branding may work as a barrier to entry by making it difficult for new firms to enter a market that is dominated by a successful brand. Note that branding need not lead to a monopoly (it is a method used also by firms in monopolistic competition and oligopoly), but it does have the effect of limiting the number of new competitor firms that enter a market. Examples of branding include brand-name items (such as NIKE®, Adidas®, Coca-Cola®, etc.).

Legal barriers

Examples of legal barriers include the following:

- **Patents** are rights given by the government to a firm that has developed a new product or invention to be its sole producer for a specified period of time. For that period, the firm producing the patented product has a monopoly on production and sale of the product. Examples include patents on new pharmaceutical products, and Intel and microprocessor chips used by IBM computers.
- **Licences** are granted by governments for particular professions or particular industries. Licences may be required, for example, to operate radio or television stations, or to enter a particular profession (such as medicine, dentistry, architecture, law and others). Such licences do not usually result in a monopoly, but they do have the impact of limiting competition.
- **Copyrights** guarantee that an author (or an author's appointed person) has the sole rights to print, publish and sell copyrighted works.
- **Tariffs, quotas and other trade restrictions** limit the quantities of a good that can be imported into a country, thus reducing competition.

Not all of these legal barriers lead to monopoly, but they all have the effect of limiting competition, thus contributing to the creation of some degree of market power.

Control of essential resources

Monopolies can arise from ownership or control of an essential resource. A classic example of an international monopoly is DeBeers, the South African diamond firm, whose control of the diamond industry peaked at 90% in the 1980s, allowing it to have a significant control over the price of diamonds. (More recently DeBeer's no longer has control and diamond prices are driven by the market.) On a national level, an example is Alcoa (the Aluminum Company of America), which, following the expiration of patents in 1909, was able to maintain its monopoly position on the production of aluminium within the United States until the Second World War, because of its control of almost all the bauxite resources within the country. On a local level, professional sports leagues create a local monopoly by signing long-term contracts with the best players and securing exclusive use of sports stadiums. A local monopoly is a single producer/supplier within a particular geographical area. Local monopolies appear more commonly than national or international ones. For example, a local grocery store in a residential area located some distance from any other stores may be a local monopoly.

Aggressive tactics

If a monopolist is confronted with the possibility of a new entrant into the industry, it can create entry barriers by cutting its price, advertising aggressively, threatening a takeover of the potential entrant, or any other behaviour that can dissuade a new firm from entering the market.

Demand and revenue curves under monopoly

The demand curve (*AR* curve) facing the monopolist

Since the pure monopolist is the entire industry, the demand or *AR* curve it faces is the industry or market demand curve, which is downward-sloping. The two demand curves shown in Figure 7.12 indicate that the perfectly competitive firm is a price-taker with zero market power, while the monopolist is a **price-maker** with a significant degree of market power.

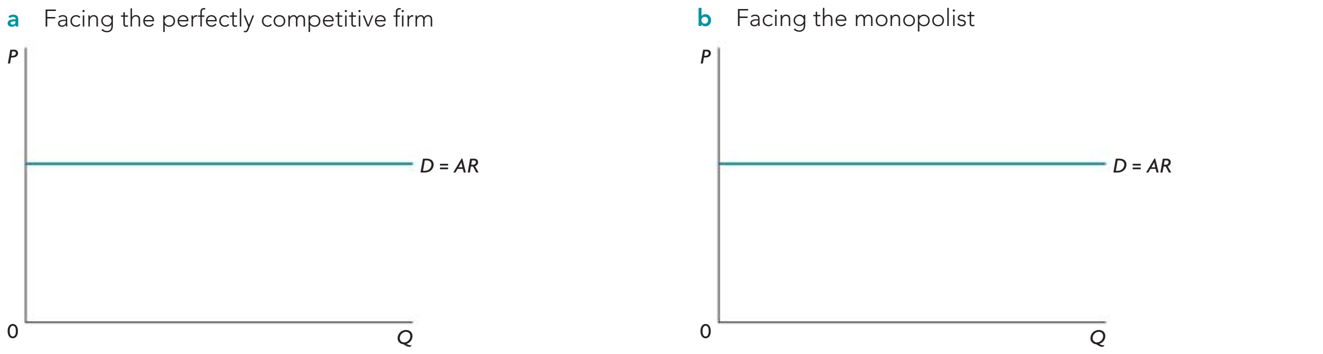


Figure 7.12: Demand curves

All firms under market structures other than perfect competition are to varying degrees price-makers, as they all face downward-sloping demand curves. Of these, the monopolist has the greatest degree of market power, or the ability to control price, because it is the sole or dominant firm in the industry.

> Market power arises whenever a firm faces a downward-sloping demand curve. Firms in all market structures except perfect competition face a downward-sloping demand curve, and therefore have varying degrees of market power, or the ability to control the price at which they sell their output; they are therefore *price-makers*.

However, whereas the monopolist has a large control over price, this control is limited by the position of the market demand curve. In Figure 7.12(b), when the monopolist chooses how much output to produce, say Q_1, it simultaneously determines the price at which the good can be sold, or P_1. It could not sell output Q_1 at a price such as P_2, since the price–quantity combination P_2 and Q_1 is at point a, lying off the demand curve.

The monopolist's revenue curves

When a firm faces a downward-sloping demand curve, price is no longer constant for all output: more output can only be sold at a lower price. Consider the example used in Table 7.3 and Figure 7.2 showing revenue data and curves for the firm that has some ability to control price. Table 7.6 provides the same data for a monopolist's total, marginal and average revenues, and Figure 7.13 plots these data.

1 Units of output (Q)	2 Product price (P) (€)	3 Total revenue TR = P × Q (€)	4 Marginal revenue $MR = \frac{\Delta TR}{\Delta Q}$ (€)	5 Average revenue $AR = \frac{TR}{Q}$ (€)
0	–	–	–	–
1	12	12	12	12
2	11	22	10	11
3	10	30	8	10
4	9	36	6	9
5	8	40	4	8
6	7	42	2	7
7	6	42	0	6
8	5	40	–2	5
9	4	36	–4	4
10	3	30	–6	3

Table 7.6: Total, marginal and average revenue when price varies with output

Looking at Table 7.6 and Figure 7.13, we may note the following:

- As price (P) falls, output (Q) increases giving rise to the downward-sloping demand curve.
- Marginal revenue, showing the change in total revenue resulting from a change in output, falls continuously; MR is equal to zero when total revenue is at its maximum (at seven units of output), and becomes negative when total revenue falls. It may be noted that the MR curve shows the slope (gradient) of the TR curve. Therefore when TR is maximum, $MR = 0$. When TR is falling, MR is negative.
- Average revenue (column 5 of Table 7.6) is equal to price (see column 2).

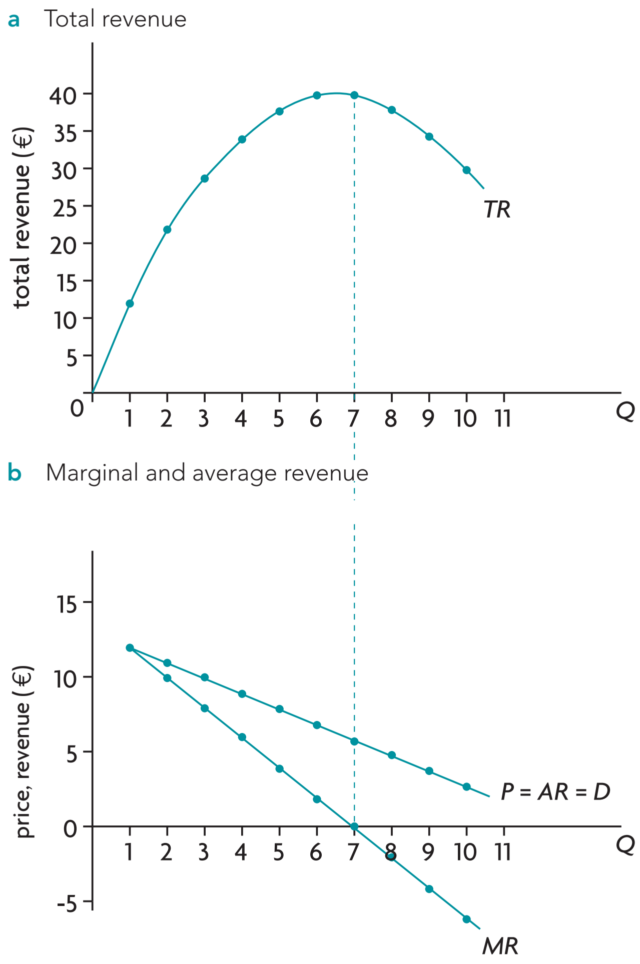


Figure 7.13: Revenue curves in monopoly

Since $TR = P \times Q$, and $AR = \frac{TR}{Q}$, it follows that *P is equal to AR.*

- *The AR and P curves represent the demand curve facing the firm.*
- *The MR curve lies below the demand curve.* The reason is that, unlike in perfect competition, where $MR = P$, here the firm must lower its price in order to sell more output. The lower price is charged not only for the last unit of output but all the previous units of output sold. Marginal revenue, or the extra revenue from selling an additional unit of output, is therefore equal to the amount of the price of the last unit sold minus what is lost by selling all the other units of output at the now lower price.[1]

TEST YOUR UNDERSTANDING 7.9

1. Outline the assumptions defining the market model of monopoly.
2. Using a diagram, explain how economies of scale can result in a monopoly market structure by posing barriers to entry.
3. Using examples, explain how branding and legal factors provide barriers to entry into an industry.
4. a Compare and contrast the demand curve facing the perfectly competitive firm and that facing the monopolist.

   b Explain why one firm is a price-taker and the other a price-maker.
5. a Explain the relationship between the monopolist's average revenue (*AR*) and marginal revenue (*MR*).

   Outline how they are related to

   b product price, and

   c the demand curve facing the firm.

Profit maximisation by the monopolist

Profit maximisation based on the marginal revenue and cost approach

The monopolist interested in maximising profit (or minimising loss) follows the same three-step approach used by the perfectly competitive firm:

i The monopolist determines the profit-maximising (or loss-minimising) level output using the $MC = MR$ rule.

ii For that level of output, it determines profit per unit or loss per unit by using

$$\frac{profit}{Q} = P - AC$$

If $AR > AC$ (or $P > AC$), the monopolist is making abnormal profit;

[1] To understand this, consider the following numerical example. Say output increases from 3 to 4 units. Marginal revenue will be the result of a gain and a loss. The gain is €9, obtained from selling the fourth unit of output at the price of €9. The loss is equal to €1 for each of the initial 3 units of output that previously were selling for €10 and must now sell for €9, equal to €3. Marginal revenue is equal to the gain minus the loss, or 9 – 3 = 6.

if AR = AC (or P = AC) it is earning normal profit (zero profit);

if $AR < AC$ (or P < AC) it is making a loss.

iii The firm multiplies $\frac{profit}{Q}$ by Q to determine total profit, or $\frac{loss}{Q}$ by Q to determine total loss.

Figure 7.14 (a), (b) and (c) show the standard AC and MC curves; these are the same as those used for profit maximisation in perfect competition. On these cost curves, the monopolist's demand and marginal revenue curves are added. Consider first part (a).

- We first find where $MR = MC$, which determines the profit-maximising level of output, Q_e.
- At Q_e, we draw a vertical line upward to the AR (or demand) curve and from there extend a horizontal line leftward to the vertical axis; this will determine the price, P_e, at which the monopolist sells output Q_e.
- For output Q_e, we find profit per unit ($\frac{profit}{Q}$), given by $P - AC$; this is the vertical distance between the average revenue (demand) and AC curves.
- To find total profit, we multiply profit per unit times the total number of units produced, which is

  $$\text{profit} = \frac{profit}{Q} \times Q$$

  and is represented by the shaded area.

The monopolist need not always make profits; it may make losses if price cannot cover AC. This is shown in Figure 7.14(c), where the monopolist is minimising loss. At the level of output Q_e, determined by $MR = MC$, the monopolist's loss is minimised. The price that will be charged is given by P_e, found by extending a line upward to the demand curve at output level Q_e. Loss per unit of output ($\frac{loss}{Q}$) is given by $AC - P$, and total loss is given by the shaded area, found by multiplying loss per unit of output by the total number of units produced.

It is also possible that the monopolist earns normal profits. This is shown in Figure 7.14(b), where Q_e occurs where $P = AC$.

Note that the distinction between the short run and the long run is not important in monopoly as it is in perfect competition. In perfect competition, this distinction is crucial because as firms enter and exit an industry in the long run, abnormal profits and losses disappear, and firms are left with normal profits in long-run equilibrium. This is not possible in monopoly, due to the presence of barriers to entry.

> Under monopoly, high barriers to entry prevent potential competitor firms from entering a profit-making industry, and the monopolist can therefore continue making abnormal profits indefinitely in the long run.

Natural monopoly

A **natural monopoly** is a firm that has economies of scale so large that it is possible for the single firm alone to supply the entire market at a lower average cost than two or more firms.

A natural monopoly is illustrated in Figure 7.15. The natural monopolist's demand curve intersects its $LRAC$ curve at a point where average costs are still falling. At the level of output Q^*, there are still economies of scale. The figure shows that this firm cannot produce any output greater than Q^* and not make a loss. If it did, then P which is given by the demand curve would be lower than $LRAC$. But as we know, $P<AC$ means the firm is making a loss. So if the firm is to earn normal profit or abnormal profit, it must produce a quantity less than or equal to Q^* and charge a price greater than or equal to AC^*.

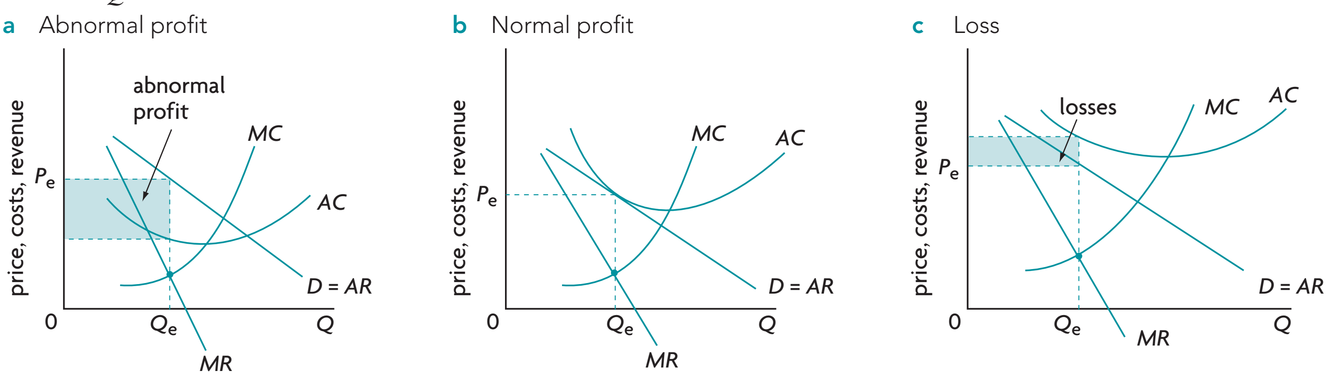


Figure 7.14: Profit maximisation and loss minimisation in monopoly: marginal revenue and cost approach

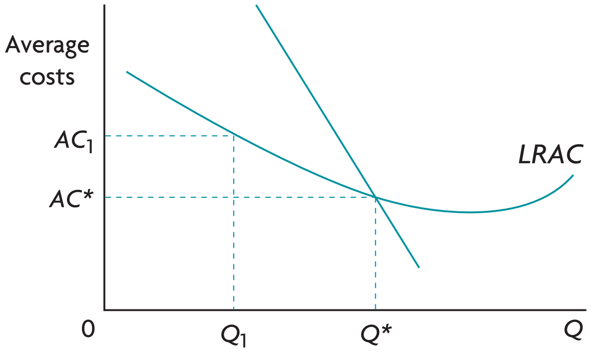


Figure 7.15: Natural monopoly

Suppose now that this firm is split into two firms of equal size. Each would produce output Q_1 and with average costs at AC_1. Clearly, average costs would be higher, and the price that would allow the firm to earn normal profit, where $P = AC$, would also be higher.

If the market demand for a product cuts *LRAC* when this is falling, this means that a single large firm can produce for the entire market at a lower average total cost than two or more smaller firms. When this occurs, the firm is called a **natural monopoly**.

Natural monopoly acts as a strong barrier to entry because potential entrants realise that it would be extremely difficult to attain the low costs of the already existing firm. High average costs would mean having to charge a high price for the product, so that the new firm would be unable to compete with the existing firm.

Examples of natural monopolies include water, gas and electricity distribution, cable television, fire protection and postal services. The falling average costs over a very large range of output often occur because of very large capital costs (such as laying pipes for water distribution, or laying cables for electricity distribution, or putting a satellite into orbit).

A natural monopoly may stop being 'natural' if changing technologies create conditions that allow new competitor firms to enter the industry and begin production at a relatively low cost. This has happened in recent years with technological change in telecommunications, forcing telephone companies that previously were natural companies to compete with new entrants into the market.

TEST YOUR UNDERSTANDING 7.10

1 Using diagrams, show the case where a monopolist
   - a earns abnormal profit (show profit per unit and total profit),
   - b earns normal profit, and
   - c incurs losses (show loss per unit and total loss).

2 a Outline the difference, if any, between the short-run and long-run equilibrium of a monopolist.
   - b Outline why a monopolist can continue to earn profits in the long run.

3 The data in the table below show the demand curve and costs (*AC* and *MC*) facing a monopolist.
   - a Calculate the monopolist's total revenue and marginal revenue for each level of output.
   - b Identify this monopolist's profit-maximising level of output.
   - c Identify the price at which this level of output be sold.
   - d Calculate the monopolist's profit per unit and total profit.

Units of output	Price ($)	Average cost ($)	Marginal cost ($)
1	10	14.0	4.0
2	9	8.5	3.0
3	8	6.3	2.0
4	7	5.0	1.0
5	6	4.4	2.0
6	5	4.2	3.0
7	4	4.1	4.0
8	3	4.3	5.0

4 a Explain the relevance of economies of scale to natural monopoly.
   - b Using a diagram and examples, explain what a natural monopoly is.
   - c Outline why a natural monopoly can be a strong barrier to entry into an industry.
   - d Suggest why governments often do not break up natural monopolies in order to increase competition.

Monopoly market outcomes and efficiency

Higher price and lower output by the monopolist compared to the industry in perfect competition

A comparison of monopoly with perfect competition at the level of the industry reveals that price is higher and quantity of output produced lower in monopoly. Figure 7.16 shows the long-run equilibrium positions of a perfectly competitive industry, composed of many small firms, and of a monopoly, which is the entire industry. Part (a) for the perfectly competitive industry shows equilibrium price and quantity to be P_{pc} and Q_{pc}, Point a, where the industry demand and supply curves intersect, appears also in part (b), showing what would happen to price and quantity if the perfectly competitive industry were organised as a monopoly. The *MC* curve of part (a), or the competitive industry's supply curve becomes the monopolist's marginal cost curve.[2] The demand curve remains unchanged, but the monopolist's marginal revenue (MR_m) curve lies below *D*. When the profit-maximising monopolist applies the $MR = MC$, the result is output Q_m and price P_m.

Since $Q_m < Q_{pc}$, the industry under monopoly produces a smaller quantity of output than the industry under perfect competition. And since $P_m > P_{pc}$, the monopolist sells output at a higher price than the perfectly competitive industry. Higher prices and lower output go against consumers' interests.

Allocative inefficiency and market failure

Loss of consumer and producer surplus

The higher price and lower output of the monopolist have important implications for consumer and producer surplus. Whereas the perfectly competitive industry achieves allocative efficiency shown by $MB = MC$ and maximum social surplus, monopoly does not. This can be seen in Figure 7.17, which is the same as Figure 7.16, only consumer and producer surplus have been added. In part (a), area A represents consumer surplus, while area B is producer surplus, with A + B showing maximum social surplus. Part (b) shows the inefficiencies that result in monopoly.

- Area C, consumer surplus in monopoly, is smaller than area A in perfect competition. Part of A was converted into producer surplus because of the higher monopoly price (P_m rather than P_{pc}), and another part of A was lost as triangle E because of the lower monopoly quantity (Q_m rather than Q_{pc}). Area E represents a welfare loss.
- Area D, producer surplus in monopoly, shows that producer surplus has increased by taking away a portion of consumer surplus (due to the monopolist's higher price), and it has also decreased by losing area F (due to the monopolist's lower quantity). Area F is also a welfare loss.
- E + F = welfare loss represents loss of social benefits (consumer and producer surplus) due to monopoly's higher price and lower quantity.

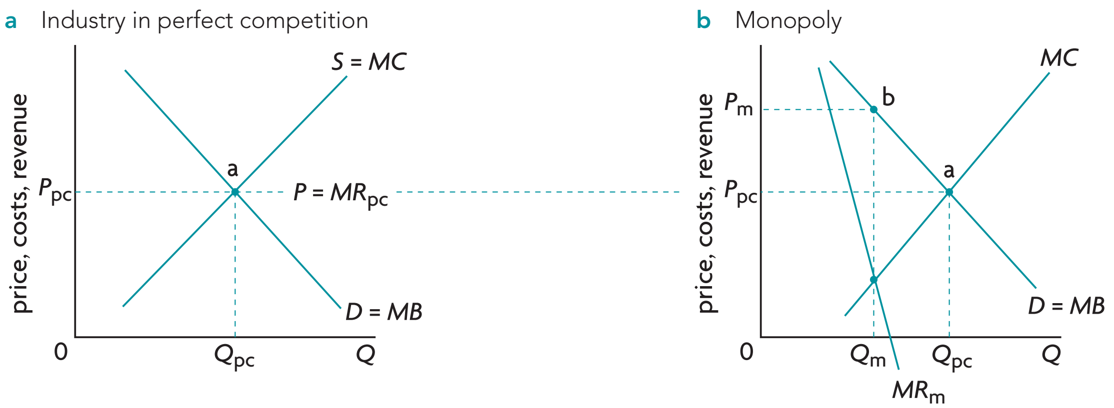


Figure 7.16: Higher price, lower output by the firm in monopoly

[2] However, note that the monopolist's *MC* curve is not its supply curve. In fact, the monopolist does not have a supply curve, because there is no single relationship between price and quantity supplied in monopoly. The reason is that the *MC* = *MR* rule for the monopolist may result in different prices for the same quantity, depending on demand conditions for the monopolist's product.

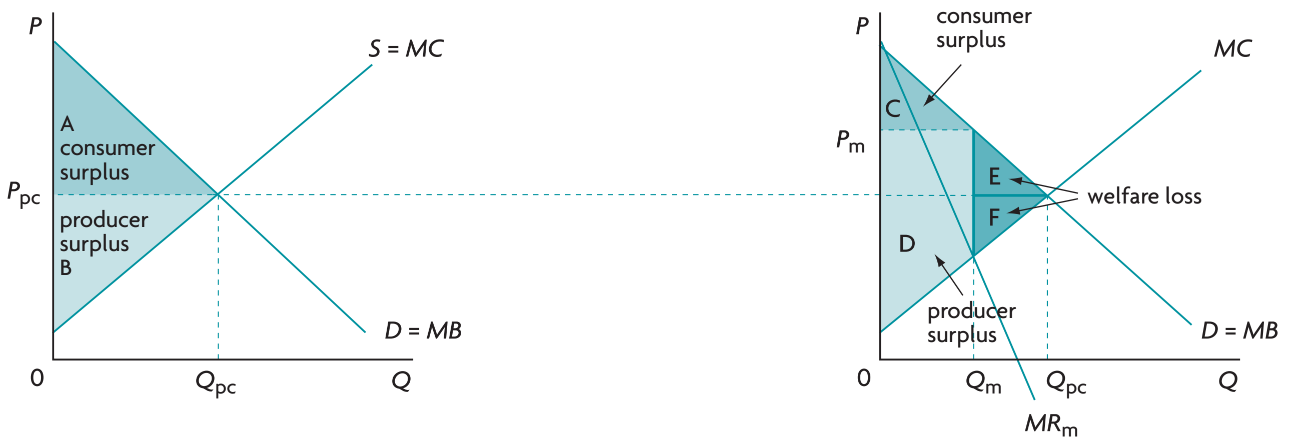


Figure 7.17: Consumer and producer surplus and welfare loss in monopoly compared with perfect competition

Note that the presence of welfare loss means that *MC* and *MB* are no longer equal. At the point of monopoly production, Q_m, $MB > MC$, meaning that there is an underallocation of resources to the good, and consumers are not getting as much of it as they would have liked.

> The presence of welfare loss in monopoly indicates market failure: there is allocative inefficiency, shown also by $MB > MC$ at Q_m, meaning there is underallocation of resources: too little of the good is produced. Also, the monopolist gains at the expense of consumers as a portion of consumer surplus is converted into producer surplus.

Allocative inefficiency: P > MC

Figure 7.18 shows the long-run equilibrium position of the firm in perfect competition and monopoly. The condition for allocative efficiency is given by $P = MC$ at the profit-maximising level of output. As we know, this condition holds for the firm in perfect competition. In Figure 7.18(b) we can see that at the profit-maximising level of output Q_m, the monopolist's price, P_e, is greater than marginal cost. This is hardly surprising, since $P > MC$ is the same as $MB > MC$ (since $P = MB$), which we saw in Figure 7.17(b).

Therefore, we conclude once again that the monopolist does not achieve allocative efficiency.

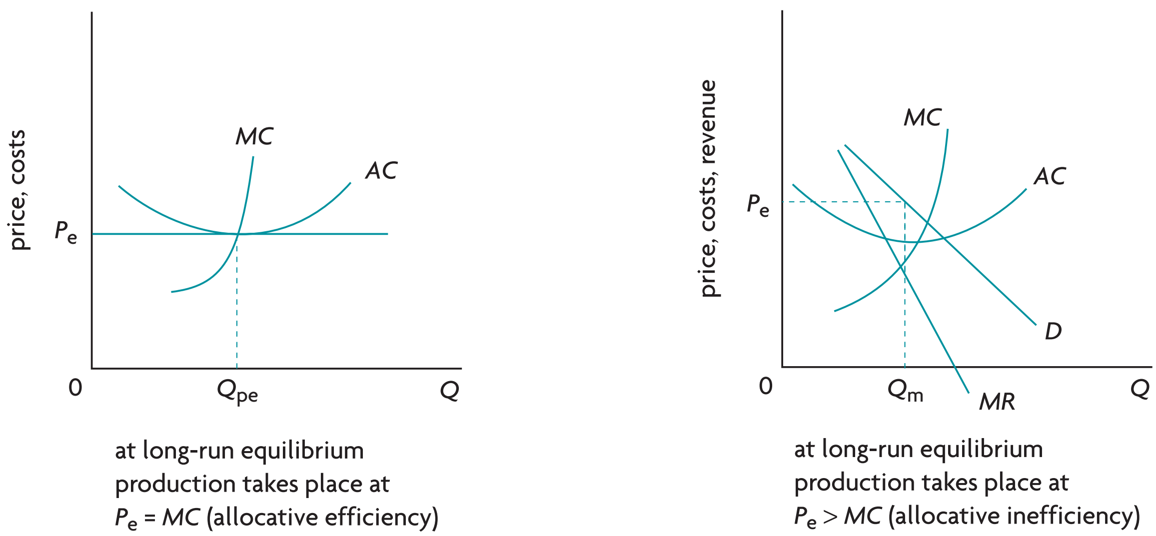


Figure 7.18: Allocative efficiency in perfect competition and allocative inefficiency in monopoly

In monopoly the underallocation of resources to the good is indicated also by $P > MC$ at the profit-maximising level of output.

Market failure and market power

The above discussion shows clearly that monopoly is a type of market failure, since it results in an underallocation of resources to the production of the good with welfare loss.

We can now also see how *market power* is connected to market failure. Earlier in this chapter market power was defined as the ability of a firm to control the price at which it sells its output. Why does this lead to market failure?

The answer lies in Figure 7.16, which compares perfect competition with monopoly. In part (a) we see the perfectly competitive firm with its horizontal demand curve accepting a P that is equal to MC, therefore there is allocative efficiency. In part (b) we see the monopolist with its downward sloping demand curve charging a P that is greater than MC, therefore there is allocative inefficiency. In fact this allocative inefficiency is the result of the downward sloping demand curve with the resulting *market power* that it provides to the firm. We can therefore redefine market power.

Market power refers to the ability of a firm to charge a price greater than marginal cost, or $P>MC$. This can only occur when a firm faces a downward sloping demand curve.

In view of their extensive market power, monopolies are held to be undesirable from a social point of view. For this reason, private, unregulated monopolies are illegal in most countries in the world.

TEST YOUR UNDERSTANDING 7.11

1. Using a diagram, compare and contrast the price and output outcomes of a perfectly competitive industry and an industry organised as a monopoly.
2. Using diagrams, compare and contrast welfare outcomes of a perfectly competitive industry and a monopolistic industry. Explain the meaning of $MB > MC$ at the monopolist's equilibrium level of output.

CONTINUED

3. Using diagrams, show whether or not the monopolist achieves allocative efficiency, and compare with the firm in perfect competition.
4. Using diagrams, explain the relationship between market power and market failure.

Evaluating monopoly and comparing with perfect competition

Criticisms of monopoly

Welfare loss, allocative inefficiency and market failure

In contrast to perfect competition, the monopolist fails to achieve allocative efficiency. Figure 7.17 shows the loss of social surplus and the appearance of welfare loss while Figure 7.18 shows that $P>MC$. Both diagrams also show that $MB>MC$, which are indications that the monopolist underallocates resources to the production of a good. Monopoly therefore represents a form of market failure.

Higher price and lower output in monopoly

The welfare loss noted above in fact arises because the monopolist produces a smaller quantity of output and sells it at a higher price than a perfectly competitive industry as shown in Figure 7.16, once again showing that monopoly is not in the interests of consumers.

Loss of consumer surplus to the monopolist

By charging a higher price than the perfectly competitive firm, such that $P>MC$ means that a portion of consumer surplus is taken away from consumers and gained by the monopolist. Figure 7.17, shows that the monopolist gains at the expense of consumers.

Negative impacts on the distribution of income

Since monopolies charge higher prices than perfectly competitive firms, there is a redistribution of income away from consumers who must pay the higher prices and toward the owners of the monopoly in the form of higher profits.

Lack of competition may give rise to higher costs

Whereas firms in perfect competition are under constant pressure to produce with the lowest possible costs to survive, under monopoly the absence of competitor firms may result in higher average costs for many possible reasons such as a poorly motivated workforce, lack of innovation, poor management or avoidance of risk. The possibility of maintaining abnormal profits over the long run due to high barriers to entry can make the monopolist less concerned about keeping costs low. This is known as *X-inefficiency*.

Possibly less innovative

While monopolies have good reasons to pursue R&D for product development and innovation, the opposite may also occur. High barriers to entry, shielding monopolies from competition, could make them less likely to innovate than firms in monopolistic competition or oligopoly (see below) which are constantly under pressure to innovate to maintain or increase their share of sales. (See Real world focus 7.2.)

Potential benefits of monopoly

Economies of scale

Economies of scale are a major argument in favour of large firms that can achieve lower costs as they grow larger. Monopolies, because of their size, are very well placed to take advantage of economies of scale. Lower average costs provide the monopolist with the possibility of lowering prices, which could possibly approach those achieved in perfect competition. Consumers can therefore gain because lower average costs may potentially translate into lower prices, as well as increased quantity of output. Society also gains from economies of scale because the lower average costs of production mean there is less waste in the use of resources. A perfectly competitive firm, due to its very small size, cannot capture economies of scale.

Natural monopoly

In the event of a natural monopoly, there are added benefits due to the achievement of very low average costs by the single firm. See the discussion above on natural monopolies.

Research and development (R&D) for product development and technological innovation

Firms in perfect competition are unlikely to engage in R&D. They have no abnormal profits in the long-run with which they can finance R&D. They sell identical products and therefore are not interested in product development that would differentiate their products. They are unable to create barriers to entry as they are too small and so have no incentive to engage in R&D.

By contrast, a number of factors suggest that monopolies have good reasons to pursue innovation:

- Their abnormal profits provide them with the ability to finance large R&D projects.
- Protection from competition due to high barriers to entry may favour innovation and product development, by offering firms the opportunity to enjoy the profits arising from their innovative activities (new inventions, new products, new technologies, etc.). This, after all, is the rationale of awarding firms patent protection for a period of time.
- Firms may use product development and technological innovation as a means of maintaining their abnormal profits over the long term, by creating barriers to entry for new potential rivals. If a firm can develop a new product that potential rivals are unable to produce, the rivals may be less likely to try to enter the industry and compete with the innovating monopolist.

TEST YOUR UNDERSTANDING 7.12

1. Using diagrams, compare and contrast the market structures of perfect competition and monopoly, emphasising the advantages and disadvantages of each.
2. Discuss
   a. the factors that may make monopoly a desirable market structure; and
   b. why monopoly as a market structure, and market power generally, come under heavy criticism and are held to be against society's best interests.

7.5 Monopolistic competition

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain that monopolistic competition has less market power due to many substitutes compared to monopoly and illustrate with a diagram showing a more elastic demand curve compared with monopoly (AO2, AO4)
- explain profit maximisation in the short run and long run, with diagrams showing abnormal profit, normal profit, loss (AO2, AO4)
- explain allocative inefficiency and market failure in monopolistic competition (AO2)
- explain that monopolistic competition has more product variety in exchange for less efficiency (AO2)
- discuss the advantages and disadvantages of monopolistic competition (AO3)

The model of *monopolistic competition* is based on the following assumptions:

- There is a large number of firms; this is similar to perfect competition.
- There are no barriers to entry, as in perfect competition.
- There is product differentiation; unlike in perfect competition where products are identical.

Product differentiation can be achieved by:

- physical differences – products may differ in size, shape, materials, texture, taste, packaging, etc. (think, for example, of the variety of clothes, shoes, books, processed foods, furniture)
- quality differences – products can differ in quality
- location – some firms attempt to differentiate their product by locating themselves in areas that allow easy access for customers, such as hotels near airports and convenience stores in residential areas
- services – some firms offer specific services to make their products more attractive, such as home delivery, product demonstrations, free support, warranties and purchase terms
- product image – some firms attempt to create a favourable image by use of celebrity advertising or endorsements, by brand names, or attractive packaging.

Examples of monopolistically competitive industries include book publishing, clothing, shoes, processed foods of all kinds, jewellery, furniture, textiles, dry cleaners, petrol (gas) stations, restaurants.

Product differentiation and the demand and revenue curves

More elastic demand curve compared with monopoly

As the term 'monopolistic competition' suggests, this market structure combines elements of both competition and monopoly. It resembles perfect competition because there are many firms in the industry and there is freedom of entry. It is like monopoly because of product differentiation. Each firm in an industry is a 'mini-monopoly' in the specific version of the good that it produces. For example, Adidas is a monopoly in Adidas® shoes, NIKE is a monopoly in NIKE® shoes, and Puma is a monopoly in Puma® shoes. This means that each of these producers faces a downward-sloping demand curve for its product. However, because each of these products is at the same time a substitute for the other, this demand curve is relatively elastic, i.e. it is more elastic than in monopoly, but less elastic than in perfect competition, as shown in Figure 7.19.

In perfect competition, if a firm raises its price, it loses all its sales to its competitors (Figure 7.19(a)). In monopoly, if a firm raises its price, it loses some but not all sales, as it is the sole producer of the good and consumers have no alternative product they can buy (Figure 7.19(b)). In monopolistic competition (Figure 7.19 (c)), if a firm raises its price, it will lose more sales than the monopolist, because consumers now do have substitutes they can switch to; but it will lose fewer sales than the perfectly competitive firm because of product differentiation – the available substitutes are not perfect substitutes, as they are in perfect competition.

This has important implications: it means that if consumers can be convinced that the product they are purchasing (for example, Puma® shoes) is superior to the available

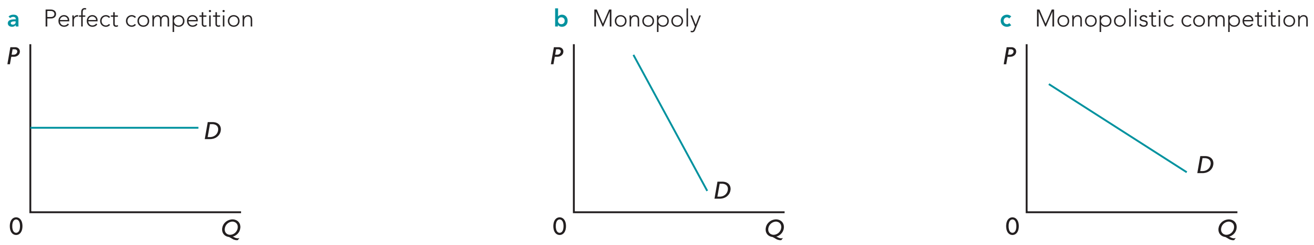


Figure 7.19: Demand curves facing the firm under three market structures

substitutes (Adidas® and NIKE® shoes), then Puma has succeeded in establishing a mini-monopoly for its product. Therefore if the price of Puma shoes increases, only some, and not all, buyers of Puma shoes will switch to other brands. Those who believe that Puma shoes are superior will continue to buy them, in spite of the higher price.

> Firms in monopolistic competition face a demand curve that is less elastic than in perfect competition but more elastic than in monopoly.

The roles of price and non-price competition

Price competition occurs when a firm lowers its price to attract customers away from rival firms, thus increasing sales at the expense of other firms. **Non-price competition** occurs when firms use methods other than price reductions to attract customers from rivals. The most common forms of non-price competition are product differentiation (including all the features noted above, such as physical and quality differences, packaging, services provision, location, etc.), advertising and branding (creating brand names for products).

Monopolistically competitive firms compete with each other on the basis of both price and non-price competition. They engage heavily in product differentiation through R&D in product development, as well as in advertising and branding. Firms that can attract customers by use of these methods increase their market power and their ability to charge a higher price without risking loss of buyers to rival firms.

In general, the more differentiated the product is from its substitutes and the more successful the advertising and branding as methods of convincing consumers about the superiority of a product, the less elastic will be the demand curve facing the firm,[3] the greater the market power (the ability to control price), and the larger the firm's potential to increase short-run profits. By contrast, firms that are less able to achieve consumer loyalty for their product, and whose product is less differentiated from substitutes, may have to rely more on price competition to increase their sales.

TEST YOUR UNDERSTANDING 7.13

1. Outline the assumptions defining the market model of monopolistic competition.
2. List some examples of monopolistically competitive firms in your neighbourhood. Analyse what makes them so.
3. Describe how a monopolistically competitive firm is like a firm in perfect competition; how it is like a monopoly.
4. Explain
   a. what firms in monopolistic competition try to achieve through product differentiation, advertising and branding, and
   b. how these activities affect the demand and revenue curves facing the firm.
5. Explain why we never see price competition and non-price competition in
   a. perfectly competitive firms, and
   b. monopolies.
   c. Explain their importance in monopolistic competition.

[3] Advertising and branding work by making the demand curve shift to the right and making it rotate so it becomes steeper. These two changes mean that demand increases and it becomes less elastic.

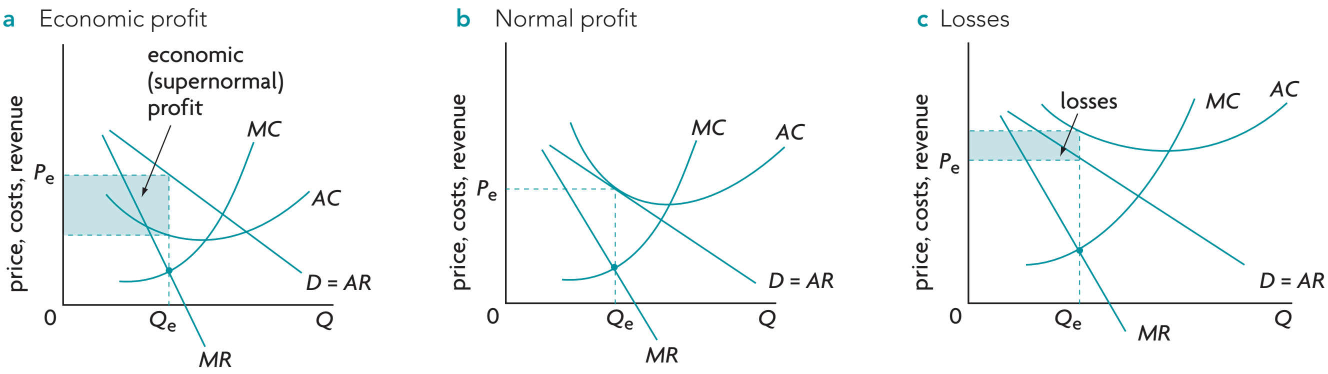


Figure 7.20: Short-run equilibrium positions of the firm in monopolistic competition

Profit maximisation

Abnormal profit and normal profit or loss in the short run

The short-run equilibrium position of the individual firm in monopolistic competition is identical to that of the monopolist, the only difference being that the demand curve is more elastic and flatter in monopolistic competition than in monopoly. In the short run, the firm can make either abnormal profit (i.e. positive profit), normal profit or losses (negative economic profit). Each of these possibilities is shown in Figure 7.20. The firm applies the $MR = MC$ rule to find the profit-maximising or loss-minimising level of output (Q_e), and then for that level of output compares price (given by the demand curve) with AC to determine profit per unit or loss per unit.

In part (a) of Figure 7.20 the firm earns abnormal profits, since $P>AC$ at Q_e; in part (b) the firm's profit is exactly zero since $P = AC$ at Q_e, and therefore the firm is earning normal profit; and in part (c), the firm is making losses because $P<AC$ at Q_e. Total profit or total loss is found by multiplying profit:

$$\frac{profit}{Q} \text{ by } Q \text{ or } \frac{loss}{Q} \text{ by } Q.$$

Normal profit in the long run

The assumption of free entry of firms in the industry is very important in determining the long-run equilibrium of the firm (like in perfect competition).

In monopolistic competition, in the long run, profit-making industries attract new entrants; in loss-making industries, some firms shut down and exit the industry. The process of entry and exit of firms in the long run ensures that economic profit or loss is zero and all firms earn normal profit.

Figure 7.21 shows the long-run equilibrium of the monopolistically competitive firm. At the level of output where $MR = MC$, $P = AC$; therefore, profit is zero and each firm is earning normal profit. This figure is the same as Figure 7.20 where it happens that the firm is earning normal profit in the short run.

The process of adjustment to normal profit in the long run is different to that of perfect competition. In monopolistic competition suppose firms are making abnormal profits in the short run. New firms will enter and they will attract customers away from the existing firms. The result will be to decrease the demand facing existing firms, so that it

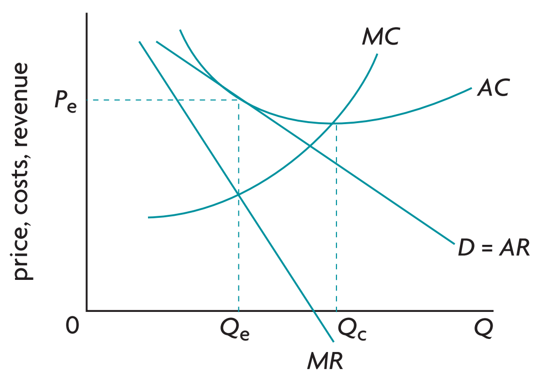


Figure 7.21: Long-run equilibrium of the firm in monopolistic competition

shifts to the left until it is just tangent to (it just touches) the AC curve. When this happens, the firms earn normal profit since $P = AC$. This is shown in Figure 7.21

On the other hand if firms are making losses in the short run, some of them will leave the industry in the long run. As they do so, their customers switch their purchases to the remaining firms, which experience an increase in demand for their product. This appears as a rightward shift of the demand curve facing them, and this process continues until losses disappear and firms are earning normal profit where $P = MC$.

Monopolistic competition, allocative efficiency and market failure

Allocative efficiency is given by the condition $P = MC$. Figure 7.21, showing the long-run equilibrium of the firm in monopolistic competition, indicates that allocative efficiency is not achieved.

Comparing price with marginal cost along the vertical line at the equilibrium level of output, Q_e, we can see that price is higher than MC, indicating that there is an underallocation of resources to the production of the good: society would have liked to have more units of the good produced.

> In monopolistic competition $P>MC$ or alternatively $MB>MC$, indicating there is market failure. The market underallocates resources to the production of the good and too little of it is produced.

Inefficiency and product differentiation

Economists argue that some inefficiency in monopolistic competition is justified by the presence of product differentiation, which leads to greater product variety. In other words consumers gain product variety while giving up some efficiency. Perfect competition, by contrast achieves efficiency but at the cost of zero product variety. Because consumers enjoy product variety, it is often argued that with monopolistic competition some inefficiency may be the 'price' consumers pay for having greater product variety.

On the other hand there is much debate among economists about how much product variety is socially desirable. Some economists argue that there is too much product variety, such as for example too many brands of breakfast cereals, detergents, cars or virtually any other product. This leads to confusion among consumers, as well as high costs of advertising and new product development which result in higher prices. It is sometimes argued that consumers would be better off with less product variety and lower prices.

Comparison of monopolistic competition with other market structures

Below is a summary of the important differences between the market models.

Monopolistic competition and perfect competition

Similarities

- Large number of firms.
- Free entry of firms into an industry (no barriers to entry).
- Normal profit in the long run, abnormal profit or loss in the short run (due to no barriers).

Differences

- **Market power and the demand curve.** Firms in perfect competition have no market power; they are price-takers facing a perfectly elastic (horizontal) demand curve they face. Firms in monopolistic competition do have some market power (ability to influence price), reflected in their downward-sloping demand curve.
- **Allocative efficiency.** Whereas the perfectly competitive firm achieves allocative efficiency in long-run equilibrium, the monopolistically competitive firm does not. Fewer than optimal resources are allocated to the production of the good. Since $P>MC$, consumers pay a higher price for the good than in perfect competition. Monopolistic competition is therefore a type of market failure.
- **Product variety.** Whereas all firms in perfect competition produce the identical product, under monopolistic competition firms go to great lengths to differentiate their products. From the consumer's perspective, product variety is usually an advantage; perfect competition cannot offer this advantage. Some inefficiency in monopolistic competition may be the 'price' consumers pay for greater product variety.
- **Economies of scale.** Firms in perfect competition cannot achieve economies of scale because they are very small. Firms in monopolistic competition may have some small room for achieving economies of scale but only to a relatively small degree as these firms also tend to be relatively small.

Monopolistic competition and monopoly

Similarities

- **No allocative efficiency therefore a form of market failure.** Both these market structures face downward-sloping demand curves, and therefore both have *MR* curves that lie below the demand curve. This means that at the profit-maximising level of output (found by $MR = MC$), $P > MC$ for both (i.e. no allocative efficiency).

Differences

- **Number of producers.** While in monopolistic competition there is a large number of firms, in monopoly there is a single firm, or else the industry is dominated by one large firm.
- **Size of firms.** In monopolistic competition firms are usually small, whereas in monopoly the fact that there is a single or dominant firm suggests a very large size.
- **Barriers to entry.** Monopolistic competition is characterised by free entry whereas in monopoly there are high barriers to entry.
- **Normal and abnormal profits.** Whereas the firm under monopolistic competition earns normal profit in the long run, the monopoly can earn abnormal profits due to high barriers to entry.
- **Competition and prices.** Free entry and exit under monopolistic competition drive abnormal profits down to zero in the long run, and allow prices to be lower for the consumer than is possible under monopoly, where barriers to entry allow the firm to maintain abnormal profits over the long run. Therefore prices under monopolistic competition are likely to be lower and quantity larger than in monopoly and more in the interests of consumers.
- **Market power.** While both monopolies and firms in monopolistic competition have market power, a monopoly is likely to have more market power because there are no substitutes for the good it produces. The availability of substitutes means that consumers can switch to substitute goods, thus reducing the firm's market power. This is another reason why prices tend to be lower in monopolistic competition than in monopoly thus favouring consumers.
- **Competition and costs.** Competition between firms in monopolistic competition puts a downward pressure on costs as firms compete with each other. These competitive pressures may force less efficient firms to leave the industry. The absence of competition in monopoly does not exert such a downward pressure on costs.
- **Research and development.** The abnormal profits that monopolies can earn over the long run puts them in a better position than monopolistically competitive firms with respect to financing R&D. However, the pressures of competition faced by monopolistically competitive firms may induce them to pursue R&D for product development in order to maintain or increase their sales.
- **Economies of scale and the possibility of natural monopoly.** Some small economies of scale may be achieved by firms under monopolistic competition, but the potential for this is much greater under monopoly, which can benefit consumers through lower prices.
- **Product variety.** Whereas many monopolies sell more than one product, there is likely to be far greater product variety in monopolistic competition which is characterised by many firms producing products that are substitutes for each other.

TEST YOUR UNDERSTANDING 7.14

1 Use diagrams to explain how a firm in monopolistic competition can
   a earn abnormal profit (show profit per unit and total profit),
   b earn normal profit, and
   c incur losses (show loss per unit and total loss).

2 Outline the role of free entry of firms in monopolistic competition in the adjustment from short-run to long-run equilibrium.

3 a Using a diagram, show the firm's long-run equilibrium position in monopolistic competition.
   b Comment on whether the firm achieves abnormal profit or normal profit in long-run equilibrium.

CONTINUED

4 Explain why the firm in monopolistic competition represents a type of market failure.

5 Explain what is meant by the idea that some inefficiency is the 'price' of product variety.

6 Evaluate monopolistic competition in comparison with

a perfect competition, and

b monopoly.

7.6 Oligopoly

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between collusive and non-collusive oligopoly and draw a diagram showing collusive oligopoly (AO2, AO4)
- explain features of oligopoly including interdependence, risk of price war, and incentive to collude versus incentive to cheat (AO2)
- explain the relevance of price and non-price competition for firms in oligopoly (AO2)
- explain the presence of allocative inefficiency and market failure (AO2)
- explain the simple game theory payoff matrix (AO2)
- explain the meaning of market concentration and concentration ratios (AO2)
- discuss advantages and disadvantages of oligopoly (AO3)

Oligopolies have the following characteristics:

- **There is a small number of large firms.** The term 'oligopoly' derives from the Greek word ολιγοπώλιο (oligopolio) meaning 'few sellers'. Oligopolistic industries are dominated by a small number of large firms, though in any one industry the firms are likely to vary in size.
- **There are high barriers to entry.** All the barriers to entry discussed under monopoly are relevant to oligopoly. They include economies of scale (for example, the aircraft and car industries); legal barriers such as patents (the pharmaceutical industry); control of natural resources (such as oil, copper, silver); aggressive tactics such as advertising or threats of takeovers of potential new firms. An additional barrier to entry in oligopoly involves high start-up costs (the costs of starting a new firm) associated with developing a new or differentiated product. Many established oligopolies spend enormous sums on product differentiation and advertising, making it difficult for new firms to match such expenditures.
- **There is interdependence.** Firms in perfect and monopolistic competition, due to their large numbers in an industry, behave independently of each other, so when they make decisions such as how much to produce they do not take the possible actions of other firms into consideration. By contrast, the small number of firms in oligopolistic industries makes the firms mutually interdependent; decisions taken by one firm affect other firms in the industry. If any one firm changes its behaviour, this can have a major impact on the demand curve facing the other firms. Therefore, firms are keenly aware of the actions of their rivals.

Most products of oligopolistic firms are differentiated such as pharmaceuticals, cars, aircraft, breakfast cereals, cigarettes, refrigerators and freezers, cameras, tyres, bicycles, motorcycles, soaps, detergents. Homogeneous or undifferentiated products include oil, steel, aluminium, copper, cement.

Interdependence, game theory and price and non-price competition

Interdependence

The interdependence of oligopolistic firms has important implications for their behaviour:

- **Strategic behaviour.** Strategic behaviour is based on plans of action that take into account rivals' possible courses of action. It is similar to playing a card game, or chess, where individual players' actions are based on the expected actions and reactions of their rival(s). Strategic behaviour of oligopolistic firms is the result of their interdependence. For example, a firm plans a course of action X assuming its rivals will follow one policy, and it plans course of action Y assuming its rivals follow a different policy. Under oligopoly, firms planning their strategies make great efforts to guess the actions and reactions of their rivals in order to formulate their own strategy.

- ***Conflicting incentives.*** Firms in oligopoly face incentives that conflict, or clash with each other:
    - *Incentive to collude* – the term *collusion* refers to an agreement between firms to limit competition between them, usually by fixing price and therefore lowering quantity produced. By colluding to limit competition, they reduce uncertainties resulting from not knowing how rivals will behave, and maximise profits for the industry as a whole.
    - *Incentive to compete, or to cheat in a collusive agreement* – at the same time, each firm faces an incentive to compete with its rivals in the hope that it will capture a portion of its rivals' market shares and profits, thereby increasing profits at the expense of other firms. If firms have formed a collusive agreement, they face an incentive to cheat on their 'partners' in the agreement in order to increase their profits at their expense.

Explaining oligopolistic behaviour by use of game theory: collude or compete/cheat

The characteristics of interdependence, strategic behaviour, and conflicting incentives are illustrated very effectively by **game theory**, a mathematical technique analysing the behaviour of decision-makers who are dependent on each other, and who display strategic behaviour. Game theory has become an important tool in microeconomics, and is based heavily on the work of American mathematician and economist John F. Nash (the subject of the 2001 film, *A Beautiful Mind*), who together with John Harsanyi and Reinhard Selten, received the 1994 Nobel Prize in Economics.

The game we will use here illustrates the *prisoner's dilemma*, showing how two rational decision-makers, who use strategic behaviour to maximise profits by trying to guess the rival's behaviour, may end up being collectively worse off. The final position that results from the game is called a *Nash equilibrium*.

Suppose there are two oligopolistic firms in the space travel industry: Intergalactic Space Travel (IST) and Universal Space Line (USL). Each firm must decide on a pricing strategy, i.e. what price to charge consumers for its space travel services, and can choose either a high-price or a low-price strategy. Each firm is interested in making its own profit as large as possible, but its profit will depend on the particular combination of pricing strategies that the two firms choose.

Figure 7.22 shows four possible combinations of pricing strategies and their corresponding profit outcomes (called 'payoffs') for the two firms. This figure represents a **payoff matrix**. For example, if both IST and USL choose the high-price strategy, in box 4, each will have profit of 40 million Zelninks (abbreviated as Zs). Box 3 shows the profit outcomes of differing price strategies; USL with a low-price strategy makes 70 million Zs, and IST with a high-price strategy makes 10 million Zs. The reason why the low-price firm makes much higher profits is that by charging a low price it captures a large portion of sales from its rival.

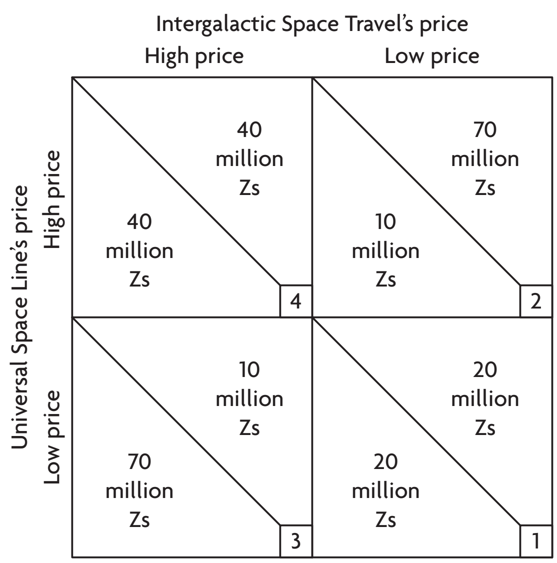


Figure 7.22: Game theory: the prisoner's dilemma

Suppose the two firms begin in box 1, where they are competing with each other on the basis of price (price competition) and therefore have a low price, leading to a low profit of 20 million Zs each. Realising that they will both be better off if they enter into a collusive agreement and charge a high price, they collaborate and agree to adopt a high-price strategy, thus entering box 4 where each one earns profits of 40 million Zs.

Now each firm faces a dilemma. Let's look at the dilemma from IST's point of view (though USL is thinking along the same lines). IST realises that by sticking to the agreement, it will continue to earn 40 million Zs, along with USL. On the other hand, IST also realises that by cheating, in other words secretly breaking the agreement, and charging a lower price, it can earn the much higher profit of 70 million Zs, while USL earns only 10 million Zs. In addition, IST realises that USL might break the agreement, in which case IST will find itself making only 10 million Zs (worse than even when it was in competition with USL, making 20 million Zs).

What should IST do? As it tries to 'outguess' USL, it is likely to cut its price to beat USL to the higher profits, but since USL is thinking along exactly the same lines, they are both likely to adopt the low-price strategy, in which case they will end up in box 1 where they both have low prices and low profits. This is the Nash equilibrium, in which both firms become worse off.

> The Nash equilibrium shows that there is sometimes a conflict between the pursuit of individual self-interest and the collective firm interest. This conflict is the prisoner's dilemma. Although the firms could be better off by cooperating, each firm, trying to make itself better off, ends up making both itself and its rival worse off.

This game illustrates many real-world aspects of oligopolistic firms, which:

- are interdependent – what happens to the profits of one firm depends on the strategies adopted by other firms; they therefore try to predict the actions of their rivals in order to plan out their own strategy
- display strategic behaviour – they plan their actions based on guesses about what their competitors are likely to do
- face conflicting incentives – they face the incentive to collude (agree to fix prices and move to box 4 where they both earn high profits); and they face incentives to compete, or in this case to 'cheat' on the agreement, by lowering their price
- become worse off as a result of price competition (trying to capture sales from their rivals by cutting prices) – since the rivals are likely to match the price cuts, all firms end up with lower prices and lower profits (box 1); this is called a **price war**
- have a strong interest in avoiding price wars, because they realise that every one will become worse off through price cutting – this creates a strong incentive for them to compete on the basis of factors other than price (non-price competition).

The roles of price and non-price competition in oligopoly

Unlike firms in monopolistic competition that compete on the basis of both price and non-price competition, oligopolistic firms go to great lengths to *avoid price competition*. They are very careful not to trigger a price war, where one firm's price cut is matched by a retaliatory price cut by another firm. As our discussion of game theory showed, a price war makes all the firms of an industry collectively worse off due to lower prices and lower profits. A price war may even lead to prices lower than average costs, and therefore losses.

However, oligopolistic firms usually do engage in intense *non-price competition*, involving efforts by firms to increase market share by methods other than price, which typically include the following:

- product development
- advertising
- branding
- numerous services such as quality customer service, warranties, provision of credit, discounts on upgrades and others.

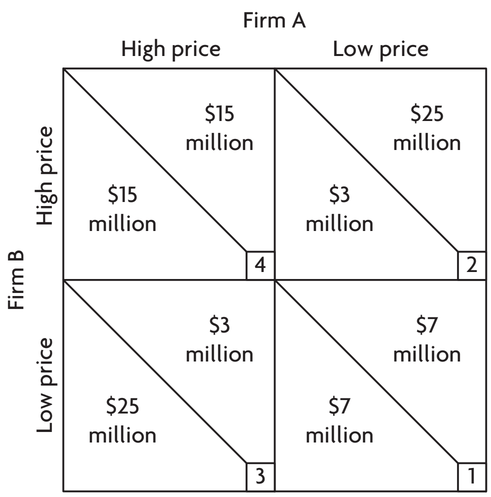
Non-price competition is very important in oligopoly for the following reasons:

- Oligopolistic firms often have considerable financial resources (due to large profits) that they can devote to both R&D and advertising and branding. Whereas monopolistically competitive firms also engage in non-price competition, their resources for these purposes are generally not as large.
- The development of new products provides firms with a competitive edge; they increase their market power, demand for the firm's product becomes less elastic, and successful products give rise to opportunities for substantially increased sales and profits.

- Product differentiation can increase a firm's profit position without creating risks for immediate retaliation by rivals. It takes time and resources for rival firms to develop new competitive products.

TEST YOUR UNDERSTANDING 7.15

1. Identify the main characteristics of an oligopoly.
2. a Identify the conflicting incentives faced by oligopolistic firms.
   b Explain how they relate to their interdependence.
3. a Referring to the conflicting incentives faced by oligopolistic firms, explain why the payoff matrix shown below illustrates the 'prisoner's dilemma' confronting the players of this game.
   b Explain the possible profit outcomes of the two firms.
   c Describe how these outcomes are related to the firms' interdependence.



4. Provide examples of non-price competition, and explain why it is important to firms in oligopoly.
5. a Outline why oligopolistic firms avoid price competition.
   b Define a price war, and outline why oligopolistic firms avoid it.

Collusive and non-collusive oligopoly

There are two predominant types of oligopolies: collusive and non-collusive.

Collusive oligopoly

Collusive oligopoly refers to situations where firms agree to *collude*, which means they form an agreement between themselves to limit competition, increase market power and increase profits. The most common form of **collusion** involves agreements to fix prices such as by holding prices constant at some level, raising prices by some fixed amount, fixing price differences between different products, and others.

Collusion is illegal in most countries, because it works to limit competition.

Cartels

A common type of collusion involves the formation of a *cartel*, which is a formal agreement between firms in an industry to take actions to limit competition in order to increase profits. The agreement may involve limiting and fixing the quantity to be produced by each firm, which results in an increase in price; fixing the price at which output can be sold; dividing the market according to geographical or other factors; or agreeing to set up barriers to entry. Whatever the case, the objective is to limit competition, increase the market power of the firms, and increase profits.

Suppose the firms of an industry decide to form a cartel by fixing price. Figure 7.23 illustrates how the cartel maximises profit. Note that this figure is identical to Figure 7.14(a), which illustrates profit maximisation for a monopolist.

The key objective of a cartel is to limit competition between the member firms and attempt to maximise joint profits. Cartel members collectively behave like a monopoly.

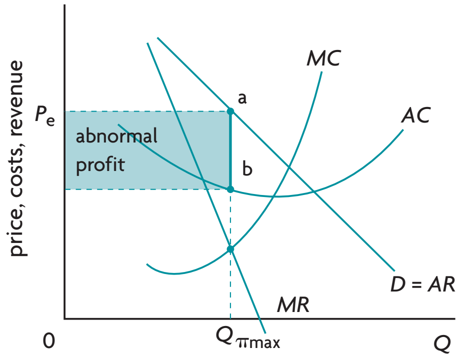


Figure 7.23: Profit maximisation by a price-fixing cartel

In Figure 7.23 the demand curve and marginal revenue curve shown are for the industry as a whole. The *MC* curve is the sum of all the *MC* curves of all the firms in the cartel. The cartel equates *MR* with *MC* to find the cartel's profit-maximising level of output, $Q_{\pi max}$, and then determines price P_e (given by the demand curve). It is then a question of dividing up industry output $Q_{\pi max}$ between all the firms, or deciding how much of the total quantity will be produced by each firm. One way this can be done is to agree on what share of the market each firm will have based on historical market shares. Another way is that firms may agree to compete with each other for market shares using non-price competition (product differentiation and advertising).

The best-known example of a cartel is OPEC (Organization of the Petroleum Exporting Countries), composed of a group of 13 oil-producing countries. OPEC periodically tries to raise the world price of oil by cutting back on its total output. Each member country is assigned an output level (quota) that it is permitted to produce. The restricted quantity of oil results in a higher price.

Firms participating in a cartel have much to gain:

- increased market power and hence the ability to control price of the product
- increased profits due to higher prices
- elimination of competition between the firms, and therefore no more uncertainty or need to outguess their rivals.

However, collusive oligopoly is illegal in most countries, for the same reasons that private unregulated monopolies are also illegal: they restrict competition, raise prices, reduce quantities of output and are held to be against consumers' and society's best interests. However, in practice it is difficult for authorities to discover and prove the existence of a cartel. It is suspected that there may be far more cases of collusive oligopoly than what are discovered.

While cartels bring important benefits to their members, they are not easy to create and maintain for several reasons.

- **The incentive to cheat.** Every firm in a cartel faces an incentive to cheat on the agreement, by offering to secretly lower the price for some buyers. But if many firms cheat, or if cheating is discovered by other firms in the cartel, then the cartel may collapse.
- **Cost differences between firms.** Since the price agreed upon by the cartel is common to all the firms, firms with higher average costs have lower profits, while lower-cost firms enjoy higher profits. Cost differences between firms lead to difficulties in agreeing on a common price.
- **Number of firms.** The larger the number of firms, the more difficult it is to arrive at an agreement regarding price and the allocation of output, as the greater number of differing views make agreement and compromise more difficult to achieve.
- **The possibility of a price war.** A possible outcome of one or more firms cheating on the cartel agreement is a price war, where one firm's price cut is matched by retaliatory price cuts by other firms. The result of a price war is to make all the firms of an industry collectively worse off due to lower prices and lower profits (in the example of game theory).

Other types of collusion

The difficulties involved in establishing and maintaining cartels as well as their illegality sometimes make firms turn towards informal types of collusion. *Informal collusion* refers to co-operation that is implicit or understood between the co-operating firms, without a formal agreement. The objectives of informal collusion are also to co-ordinate prices, avoid competitive price-cutting, limit competition, reduce uncertainties and increase profits.

One type of informal collusion is *price leadership*, where a dominant firm in the industry (which may be the largest, or the one with lowest costs) sets a price and also initiates any price changes. The remaining firms in the industry become price-takers, accepting the price that has been established by the leader. The implicit agreement (as there is no formal agreement) binds the firms as far as price goes, but they are free to engage in non-price competition. A characteristic of price leadership arrangements is that price changes tend to be infrequent, and are undertaken by the leader only when major demand or cost changes occur. Examples include the airline industry, supermarkets and fast food restaurants.

Non-collusive oligopoly

Non-collusive oligopoly refers to oligopolistic firms that do not collude in any way in order to fix or coordinate prices and limit competition. Each firm behaves independently; however, they are still aware of each other in their pricing decisions and display strategic behaviour in that they take the possible actions of their rivals into consideration.

It can be observed that in the real world, prices of oligopolistic industries tend to be rigid or inflexible; once a particular price is reached, it tends to be relatively stable over long periods of time. Moreover, in situations when prices do change, they tend to change together for all the firms in an industry. Such price rigidities can be easily explained by collusive oligopoly, since firms behave like a monopoly in such cases and so tend to have similar prices, but how can they be explained in situations where firms do not collude?

We can understand the observed price stability of oligopolistic firms that do not collude using the idea of strategic behaviour. Suppose there are three oligopolistic firms, A, B and C producing a similar product. A, B and C do not collude or communicate with each other in any way, instead their pricing behaviour is strategic, and is strongly influenced by their expectations of how rival firms will react if they undertake a price change.

Suppose firm A considers a price change, but before changing (increasing or decreasing) its price, it tries to predict how firms B and C will react, and what will be the consequences of their reaction. Firm A's reasoning is as follows:

- If I raise my price, what will B and C do? They are unlikely to increase their price, because if they continue to sell at their lower price they will take away a portion of my sales, they will be better off and I will be worse off. Therefore I should *not increase my price*.
- If I drop my price, what will B and C do? They are likely to drop their price as well, and so as a result, I am unlikely to be much better off than I am now, and I may be worse off with a lower price, so I should *not drop my price*.

This line of reasoning is the same for all three firms, A, B and C, and tends to result in price stability.

This simple idea illustrates three important points:

- **Firms that do not collude are forced to take into account the actions of their rivals in making pricing decisions.** Otherwise they risk lowering their revenues and profits, which in turn could lead to price instability.
- **Even though the firms do not collude, there is still price stability.** Firms are reluctant to change their price because of the likely actions of their rivals, which could result in lower profits for the firm initiating price changes.
- **Firms do not compete with each other on the basis of price.** They do not try to increase their sales by attracting customers through lower prices. A lower price not only invites price cuts by rivals, with resulting lower profits for all the firms, but also risks setting off a *price war*.

The concentration ratio

As oligopolies involve a small number of large firms that dominate an industry, it is important to know how 'concentrated' the industry's output is among the industry's largest firms, as this information may provide clues on whether oligopolies have too much market power. A **concentration ratio** provides an indication of the percentage of output produced by the largest firms in an industry; it measures **market concentration**. There is no fixed number of firms for which a concentration is calculated. For example, we could say that the 3-firm concentration ratio of industry X is 78%, which means that the three largest firms of industry X produce 78% of the industry's total output; or the 4-firm concentration ratio of industry Y is 45%, which means that the four largest firms in industry Y produce 45% of the industry's total output. Table 7.7 provides some examples of concentration ratios in the United States. We can see there are wide variations from industry to industry, with the most concentrated industry of those appearing in the table being transportation equipment and the least concentrated being furniture.

Industry	4-firm	8-firm
Food	35.0%	46.7%
Chemicals	44.8%	58.7%
Transportation equipment	64.3%	77.7%
Furniture	18.8%	26.5%
Textile mill products	32.0%	44.9%

Source: *http://www.swarthmore.edu/SocSci/Economics/fpryor1/Concentration_ratios.pdf*

Table 7.7: Selected concentration ratios in domestic US manufacturing

Concentration ratios provide an indication of the degree of competition in an industry. They suggest that the higher the concentration ratio, the lower the degree of competition, while a low concentration ratio would indicate a greater degree of competition.

In general, an industry is considered to be oligopolistic if the four largest firms control 40% of output. (This is an arbitrary cut-off point, as there is nothing special about a concentration ratio of 40%.)

Concentration ratios have several weaknesses that limit their usefulness as a measure of the degree of competition:

- Whereas concentration ratios reflect concentration in a national market, they do not reflect competition from abroad, arising from imports.
- Concentration ratios provide no indication of the importance of firms in the global market; there may be some competition in a domestic market, but the firms may have a very strong, or dominant position in the global market.
- Concentration ratios do not account for competition from other industries, which may be important in the case of substitute goods, such as in the case of different metals. Whereas there may be a high concentration ratio in the aluminium industry, for example, this would be lower if considered together with copper, with which aluminium competes.
- Concentration ratios do not distinguish between different possible sizes of the largest firms. For example, a three-firm concentration ratio of 90% could consist of three firms with 30% of the market each, or of three firms, one of which has 60% of the market and the other two have 15% each.

REAL WORLD FOCUS 7.1

Pharmaceutical oligopoly faces lawsuit

Forty-four states in the United States filed a lawsuit in 2019 against 20 pharmaceutical firms for fixing prices of over 100 common generic (non-patented) medicines, including treatments for diabetes and cancer. Following a five-year investigation, the firms have been accused of illegal conspiracies to either stop prices from falling or to raise prices, in some cases by more than 1000%. The Attorney General of the State of Connecticut states that:

'We have hard evidence that the generic drug industry perpetrated a multi-billion-dollar fraud on the American people. We have emails, text messages, telephone records and former company insiders that we believe will prove a multi-year conspiracy to fix prices and divide market share for huge numbers of generic drugs.'

Figure 7.24: Capsules containing business and economics formulas

The legal action seeks damages, civil penalties and court actions to restore competition to the generic drug market. Generic drugs are lower-priced alternatives to brand-name drugs and as a result save drug buyers and taxpayers billions of dollars a year. Yet according to the lawsuit:

'Prices for hundreds of generic drugs has risen – while some have skyrocketed, without explanation, sparking outrage from politicians, payers and consumers across the country.'

Sources: *www.cnbc.comw; www.bmj.com*

Applying your skills

1. Identify this pharmaceutical oligopoly as collusive or non-collusive, offering reasons why.
2. Draw a diagram and use it to explain how this oligopoly fails to achieve allocative efficiency.
3. Normally, it is very difficult to determine the presence of collusion. Outline why this may not be the case here.

Evaluating oligopoly

Oligopoly, allocative efficiency and market failure

As firms in collusive oligopoly behave like a monopoly, it is clear that there is welfare loss, allocative inefficiency and market failure, as Figure 7.22 demonstrates.

Yet the same is also true of oligopolistic firms that do not collude. They also face a downward sloping demand curve, and the market produces a level of output that is below the level where social surplus is maximum. Therefore, there is welfare loss here too as in the case of collusive oligopoly.

Criticisms of oligopoly

To the extent that oligopolistic firms succeed in avoiding price competition, they achieve a considerable degree of market power, and therefore face similar criticisms as monopoly:

- Welfare loss, allocative inefficiency and market failure.
- Higher prices and lower quantities of output than under competitive conditions.
- Loss of consumer surplus to the oligopolists due to higher prices resulting in $P>MC$.
- Negative impacts on the distribution of income.
- There may be higher production costs due to lack of price competition.
- Possibly less innovative.

In addition, there is a further argument against oligopoly:

- Whereas many countries have anti-monopoly legislation that protects against the abuse of market power, the difficulties of detecting and proving collusion among oligopolistic firms means that such firms may actually behave like monopolies by colluding and yet may get away with it.

Benefits of oligopoly

The benefits of oligopoly are also similar to the benefits of monopoly:

- Economies of scale can be achieved due to the large size of oligopolistic firms, leading to lower production costs to the benefit of society and the consumer (through lower prices).
- Product development and technological innovations can be pursued due to the high abnormal profits from which research funds can be drawn. This benefit of oligopoly is more important than in the case of monopoly, since non-price competition forces firms to be innovative in order to increase their market share and profits.
- Technological innovations that improve efficiency and lower costs of production may be passed to consumers in the form of lower prices.

Over and above the benefits of oligopoly that are similar to monopoly, oligopoly also offers the following advantage:

- Product development leads to increased product variety, thus providing consumers with greater choice (monopoly does not offer much product differentiation and variety).

TEST YOUR UNDERSTANDING 7.16

1. a. Outline the meaning of collusion.
   b. Identify what firms in an oligopoly try to achieve through collusion.
2. a. Define a cartel.
   b. Using a diagram, show how a cartel resembles a monopoly.
   c. Explain why cartels are illegal in most countries.
3. a. Outline why the cartel members face the incentive to cheat.
   b. Identify some obstacles to forming and maintaining cartels.
4. Explain how non-collusive oligopoly differs from collusive oligopoly.
5. a. Explain the meaning of 'concentration ratio' and provide examples (they may be hypothetical).
   b. State the purpose of calculating concentration ratios.
   c. Identify some shortcomings of concentration ratios.
6. 'Oligopoly is a type of market failure.' Justify this statement.
7. Discuss potential advantages and disadvantages of oligopolistic firms.

THEORY OF KNOWLEDGE 7.1

Perfect competition and the real world

The model of perfect competition is the basis of an idealised free-market economy, where price is the rationing system that supplies answers to the *what* and *how to produce* questions.[4] The market and the price mechanism are the means by which allocative efficiency is achieved, thus avoiding waste of resources.

Many economists have strongly criticised preoccupation with the perfectly competitive model. They note that the real world is dominated by large oligopolistic firms and monopolistically competitive firms, and the model of perfect competition bears no relationship to either of these. Trying to test the perfectly competitive model would be meaningless given the real-world context of firm behaviour. The very concept of 'competition' has vastly different meanings, with competition in perfect competition being anonymous and equivalent to complete absence of market power (no ability to influence price), and competition in the other two market models involving efforts to capture market shares from rival firms, either through price cuts or through product differentiation and advertising (non-price competition). These real-world types of competition are completely irrelevant for the perfectly competitive firm that can neither gain anything by lowering its price nor can it differentiate and advertise its product.

In addition, many economists question the notion of equilibrium (an idea borrowed from physics), and point out that equilibrium for the firm makes sense only under conditions of complete certainty and perfect knowledge. As there is no such thing in the real world, there can be no equilibrium.

On the other hand, defenders of the perfectly competitive market model argue that this can be used as a 'parable' that can approximate (or describe very roughly) the outcomes of real-world firm behaviour, regardless whether these firms are oligopolistic or monopolistically competitive. They also argue that in spite of its lack of realism, this model serves as a tool for assessing real-world situations that depart from the ideal of allocative efficiency, thus helping governments prescribe policy measures to deal with issues in the industrial sector.

Thinking points

- How useful do you think is the model of perfect competition?
- Do you think it matters that it is based on highly unrealistic assumptions (see also Theory of knowledge 2.3 in Chapter 2)?
- Do you think economists should focus more on developing and using more realistic market models, based on monopolistic competition and oligopoly?

[4] It also answers the *for whom to produce* question on income distribution. However, the answer provided is generally highly unsatisfactory, and for this reason becomes a normative issue about how governments should intervene in markets to change market-determined income distribution (see Chapter 12).

7.7 Government intervention in response to abuse of market power

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- discuss advantages of large firms with significant market power, including economies of scale and investments in R&D financed by abnormal profits, innovation (AO3)
- discuss risks with respect to output, price, consumer choice in markets dominated by large firms (AO3)
- discuss advantages and disadvantages of government intervention to deal with abuse of market power including legislation and regulation, government ownership, fines (AO3)

Advantages and risks of large firms with significant market power

In the discussion above on monopoly and oligopoly we saw that large firms with significant market power offer a number of potential advantages, including:

- advantages arising from large economies of scale and natural monopoly
- the ability of such firms to carry out R&D on new product and technology development.

We have also seen that there are a number of disadvantages and risks in terms of:

- allocative inefficiency and welfare loss
- higher prices and lower output
- loss of consumer surplus to the large firm or firms
- negative impact on the distribution of income
- higher than necessary average costs due to lack of price competition
- possibly less innovative.

Government intervention to address abuse of market power: an evaluation

The meaning of *abuse* of market power

We have seen that all market structures other than perfect competition represent market failure, resulting in a price above marginal cost ($P > MC$), and welfare loss to a lesser or greater degree. However **abuse of market power**, also known as *anti-competitive practices*, refers to situations where firms engage in activities that result in reduced competition. According to the European Commission (of the EU):

> *'A company can restrict competition if it is in a position of strength on a given market. A dominant position is not in itself anti-competitive, but if the company exploits this position to eliminate competition, it is considered to have* ***abused*** *it.'*

Examples include:

- charging unreasonably high prices
- depriving smaller competitors of customers by selling at artificially low prices they can't compete with
- obstructing competitors in the market (or in another related market) by forcing consumers to buy a product which is artificially related to a more popular, in-demand product
- refusing to deal with certain customers or offering special discounts to customers who buy all or most of their supplies from the dominant company
- making the sale of one product conditional on the sale of another product.[5]

In general we can say the following:

- *Monopolies* on the whole have the highest degree of market power which is abusive due to the lack of competition; this is why private, unregulated monopolies are illegal most everywhere.
- *Firms in oligopolies* may or may not abuse their market power. The market power of a collusive

[5] http://ec.europa.eu/competition/consumers/abuse_en.html

REAL WORLD FOCUS 7.2

Risks of increasing market concentration

In 2019 the International Monetary Fund (IMF, see Chapter 20) published a study providing evidence in support of the very popular notion that competition is weakening as markets are becoming more and more concentrated. The study examined markups over marginal costs (the amount added to cost to obtain the price) of 900 000 firms in 27 countries. Markups over marginal cost are an indicator of market power. It found that markups increased by 8% on average in the period 2000–2015, increasing most in the United States and to a lesser extent in Europe.

But the study also found that the rising markups are due to only a small share of the firms. The 10% of firms with the highest markups increased these by more than 30% compared to their competitor firms. Moreover, mergers and acquisitions, which clearly result in increased concentration, are followed by significantly higher markups.

Figure 7.25: David vs Goliath business competition, illustrating great difficulties for small businesses against large corporations

In addition it found that higher markups are linked with less investment in physical capital, which is estimated to have resulted in lower GDP growth by one percentage point in advanced economies. Further, the firms with the highest markups pay workers a smaller share of the value they create, contributing to income inequality. Finally, the study warns that market power may result in less innovation as firms become too comfortable in their position due to decreasing competitive pressures.

These negative trends may worsen if market power is not checked.

Source: *www.economist.com*

Applying your skills

1. Outline the meaning of increasing market concentration
2. Describe the evidence used by the IMF to conclude that there is increasing market concentration.
3. Explain the risks to consumers and the economy of increasing concentration of markets.

oligopoly is similar to that of a monopoly, and represents a form of abuse; therefore collusion is illegal most everywhere. Non-collusive oligopolies may or may not abuse their market power, depending on the situation.

- *Firms in monopolistic competition* have less market power which for the most part they do not abuse because there is significant competition among firms that produce substitute goods.

Therefore government intervention in response to abuse of market power occurs mainly in the cases of monopoly and oligopoly.

Legislation to protect competition

Most countries have laws that try to promote competition by preventing *collusion* between oligopolistic firms, as well as preventing anti-competitive behaviour by a single firm that dominates a market (see Real world focus 7.1). This is known as *competition policy*. Firms that are found guilty of anticompetitive behaviour are usually asked to pay fines (see Real world focus 7.5), or may be broken up into smaller firms.

Difficulties that may arise in connection with competition policies include:

- Possible difficulties in interpreting the legislation in connection with the behaviour of the offending firms. Different people may have different views on what actions involve anti-competitive behaviour. The laws themselves may be vague, allowing much room for different interpretations. According to the OECD, 'Determining when a firm's behaviour is an abuse of market power, as opposed to a competitive action, is one of the most complex and controversial areas in competition policy.'[6]
- Laws in a particular country may be enforced to varying degrees, with some governments enforcing them more strictly than others, depending on their priorities or their political and ideological views. Some governments may accept the principle that government intervention in the market in the form of strict enforcement of competition policies is necessary to protect consumers against monopolistic practices and to achieve allocative efficiency. Other governments may accept the principle that government intervention in the market is not necessary to achieve consumer protection and allocative efficiency, because over long periods of time, the market and competitive forces on their own accomplish these functions. There is no 'right' or 'wrong' answer to this issue, as it depends on normative ideas about the economy.
- If firms collude, it is difficult to discover evidence of the collusion and to prove it, as collusion occurs secretly, since it is illegal.

Legislation in the case of mergers

A merger is an agreement between two or more firms to join together and become a single firm. Mergers may occur for a number of reasons, such as an interest in capturing economies of scale (a single larger firm may be able to produce at lower average costs), or an interest in firm growth (the firms would like to become larger), or interest in acquiring market power, which is made possible by the larger size of the new, larger firm.

Mergers are an issue in competition policy because of the possibility that the single firm created from the merger have too much market power. Legislation usually involves limits on the size of the combined firms. (See Real world focus 7.4.)

Difficulties with merger policies include questions and uncertainties about what firms should be allowed to merge and what firms should not, related to issues of interpreting the legislation as well as ideological differences among different governments on the desirability or not of a high degree of market power.

The imposition of fines

Fines are often imposed if a government agency responsible for investigating anti-competitive behaviour discovers some wrongdoing (see Real world focus 7.5 below for examples).

A problem with fines is that firms will often get their lawyers to calculate whether breaking the law or complying with the law is more costly (or more profitable). Often, the profits of firms that abuse market power are great enough that they are better off illegally abusing their power and paying fines, than not abusing their power and not paying fines. Moreover, often they are not caught for anti-competitive behaviour, possibly because of the political climate, or because they are careful to cover their actions.

Big firms often neglect the ethics of wrongful behaviour if they believe that getting caught is not as costly as compliance.

[6] http://www.oecd.org/competition/abuse

REAL WORLD FOCUS 7.3

Conflicting views regarding regulation

In 2018 the European Commission (of the EU) imposed a fine of $5 billion on Google for violating the EU's competition rules by requiring phone manufacturers to preinstall Google Search and its browser app, Chrome, in order to access Google Play, Google's app store. According to the EU, 'These practices have denied rivals the chance to innovate and compete on the merits. They have denied European consumers the benefits of effective competition in the important mobile sphere.'[7]

Figure 7.26: European Commissioner Margrethe Vestager, 18 July 2018, when the EU gave Google 90 days to end illegal practices or face further fines (after being fined $5billion)

Soon after President Donald Trump tweeted 'The European Union just slapped a Five Billion Dollar fine on one of our great companies, Google. They truly have taken advantage of the U.S., but not for long!' Some years earlier President Barack Obama had stated:

Sometimes the European response here is more commercially driven than anything else. We have owned the internet, our companies have created it, expanded it, perfected it in ways that they [European firms] can't compete. And oftentimes what is portrayed as high-minded positions on issues sometimes is just designed to carve out some of their commercial interests.[8]

Later US Senator Elizabeth Warren argued:

Today's big tech companies have too much power—too much power over our economy, our society, and our democracy. They've bulldozed competition, used our private information for profit, and tilted the playing field against everyone else. And in the process, they have hurt small businesses and stifled innovation.[9]

To deal with this, she proposes breaking up the big tech companies.

Tyler Cowen, a US economics professor disagrees, arguing that large tech companies are major innovators, and breaking them up could weaken them and therefore hurt innovation.

Sources: *https://europa.eu; www.newyorker.com; https://medium.com/*

Applying your skills

1 Describe why the European Commission thought Google was abusing its market power.

2 To what extent do you think the European Commission's concerns are legitimate?

3 Compare and contrast the policies of an imposition of a fine versus legislation to break up Google.

[7] https://europa.eu/rapid/press-release_IP-18-4581_en.htm
[8] https://www.newyorker.com/news/our-columnists/why-did-the-european-commission-fine-google-five-billion-dollars
[9] https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c

REAL WORLD FOCUS 7.4

Competition authorities reject UK supermarket merger

In 2019, Britain's Competition and Markets Authority (CMA) blocked a merger between Asda, the United Kingdom's second largest supermarket, with Sainsbury's, the third largest. The two supermarkets combined would have had a market share of 30.7%, and would have become the largest supermarket in the country. In their proposed merger, Asda and Sainsbury's argued that the merger would have allowed them to cut costs with resulting price decreases of as much as 10%. However, the CMA claimed 'we have found this deal would lead to increased prices, reduced quality and choice of products, or a poorer shopping experience for all of their UK wshoppers.'

Figure 7.27: Flint, UK. Asda and Sainsbury's

There is growing concern in the United Kingdom that there is increasing concentration of firms in the domestic economy, where there have been \$2 trillion worth of mergers and acquisitions in the past decade. There are similar concerns in the United States that the economy has become too concentrated, limiting competition and reducing welfare (see also Real world focus 7.2). Therefore the CMA's decision on the supermarket merger 'could yet prove to be an important moment in the reform of Britain's over-concentrated economy'.

Source: *www.economist.com*

Applying your skills

1 Use a diagram to explain why the merger of the two supermarkets might have resulted in lower average costs.

2 Explain the theoretical arguments behind the CMA's reasoning that the merger would likely result in higher prices and reduced quality and choice of products.

The case of natural monopoly

While most countries around the world do not encourage monopoly, an exception is made if there is a natural monopoly, because it is not in society's interests to break it up into smaller firms, as this would result in higher average costs and a waste of resources.

Government ownership of natural monopolies

One possible solution to natural monopoly is to nationalise it, which involves transfer of ownership from the private sector to the public sector (the government). Government ownership allows the government to regulate natural monopolies, forcing them to lower prices and increase quantities produced in the interests of consumers, thereby reducing allocative inefficiency and welfare loss.

However government ownership sometimes leads to inefficiencies and higher than necessary costs of production, as governments are not driven by the goal to maximise profits. In general since the 1990s there has been a trend around to world to privatise (the opposite of nationalise) government enterprises.

Government regulation of natural monopolies

Governments usually regulate natural monopolies, to ensure more socially desirable price and quantity outcomes, and this can be done even when the monopoly remains under private ownership. There are two ways this can be done.

REAL WORLD FOCUS 7.5

Fines imposed by the European Union

In 2019 a group of banks (Barclays, J.P. Morgan, MUFG and Royal Bank of Scotland) were fined $1.2 billion for rigging the multi-trillion-dollar foreign exchange market for 11 currencies. The Swiss bank UBS was exempted from the fine because it alerted the existence of two cartels to the European Union.

At the same time another group of banks (Barclays, BNB Paribas, Citigroup, J.P. Morgan, Royal Bank of Scotland and UBS) were fined $2.8 billion by U.S. regulators also for rigging exchange rates.

Source: https://www.cnbc.com/2019/05/16/eu-regulators-fine-five-banks-for-forex-rigging.html

Figure 7.28: Mountain View, California, USA. One of the Google buildings

- In 2019, Google was fined €1.5 for restricting competition by imposing anticompetitive contractual restrictions on third-party websites for over ten years.
- In 2018, Google was fined €5 billion for requiring phone manufacturers to preinstall Google Search and its browser app, Chrome, in order to access Google Play, Google's app store.
- In 2018, Qualcomm was fined $1.2 billion for making 'significant payments' to Apple for using Qualcomm's chips rather than competing chips.
- In 2017, Google was fined €2.7 billion for giving its Google Shopping Service an illegal advantage in the search results.
- In 2009, Intel was fined $1.45 billion for offering customers price reductions if they used Intel's microprocessors instead of its rival AMD.

Sources: *www.theguardian.com; www.businessinsider.com; www.reuters.com;*

Applying your skills

Outline how each of these violations matches the European Commission's definition and examples of abuse of market power (see Section *The meaning of abuse of market power* above).

- **Marginal cost pricing**

The government can force the monopoly to charge a price equal to marginal cost, since with $P = MC$ the monopolist will achieve allocative efficiency, with P falling and Q increasing to the socially desirable level. This is called *marginal cost pricing*.

However, marginal cost pricing leads to losses for the natural monopolist. The reason is that $P = MC$ results in a price that is too low for the firm to be able to cover its average costs. As a result, the firm will either go out of business, otherwise the government would have to subsidise it in order to cover its losses. (This need not occur in a monopoly that is not a natural monopoly.) For these reasons marginal cost pricing is not a popular way to regulate natural monopolies.

(The interested reader may refer to the digital coursebook, illustrating a natural monopoly with marginal cost pricing, as Supplementary material.)

- **Average cost pricing**

To avoid creating losses for the natural monopolist, governments can force the firm to charge a price equal to its average costs, where $P = AC$, meaning that the firm earns normal profit. This is called *average cost pricing*. This results in a higher price and lower quantity than marginal cost pricing. However, it leads to a price and quantity combination that is superior to that of the unregulated monopolist, meaning that price is lower and quantity greater.

Although allocative efficiency is not achieved through average cost pricing, this policy offers two very important advantages: (a) the monopolist makes normal profit and is not in danger of having to shut down; and (b) it is more efficient than the market solution.

Yet, average cost pricing also has disadvantages. A monopolist in a free, unregulated market faces incentives to keep its average costs low, in order to maximise profits. If, through regulation, it is guaranteed a price equal to its average costs, it loses this incentive. If average costs increase due to inefficiency, it will still receive a price covering its costs.

Another possible disadvantage is that the regulated monopoly may continue to survive as a monopoly, even though it may stop being a natural monopoly (if technological improvements change cost conditions, such as in telecommunications). Continued regulation provides protection to the firm from new competitors that would have been able to produce more efficiently.

TEST YOUR UNDERSTANDING 7.17

1 Discuss
   - a potential advantages of large firms with significant market power, and
   - b potential risks in markets dominated by large firms.

2 Outline the meaning of abuse of market power.

3 Discuss advantages and disadvantages of possible government interventions to deal with abuse of market power.

4 a Using a diagram, explain the meaning of natural monopoly.
   - b Discuss policy options for governments to deal with natural monopoly.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Research and find an example of collusive oligopoly in your country of residence or in any other country of your choice. Try to discover how the firms colluded, how their collusion was discovered, what if any were the fines imposed, and what were the consequences.

2 Research a case of competition policy in a country and industry of your choice where firms have been accused of anti-competitive behaviour. Explain why the behaviour runs against the interests of consumers and society, and how the actions taken by the competition authority would correct the problem.

3 There is a major ongoing debate about market power concentration in the large tech firms like Amazon, Google and Facebook. According to many observers this is excessively high and growing. Research and explain the arguments that favour regulation and legislation to limit the market power of such firms. Explain the possible risks to consumers and society of the rising market power of these firms. Describe the various policies that are being proposed to regulate this industry.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Unit 3

Macroeconomics

Macroeconomics studies the economy as a whole. We now focus on the larger picture of the economy, composed of collections of many consumers, firms, resource owners and markets. Instead of individual product prices, we study the general price level of the economy; instead of demand for individual products, we examine total demand for goods and services; and instead of individual firm and industry supply, we examine the total output produced in the economy. Further, we study total employment, total investment, total exports and imports and more such totals or wholes, which in macroeconomics are called *aggregates*.

In our study of macroeconomics we will develop tools that will allow us to analyse a variety of policies governments can pursue in order to achieve a number of important macroeconomic objectives, such as low inflation and unemployment, economic growth and a more equitable distribution of income.

Real world issue 1: What accounts for variations in economic activity over time, and why is this important?

CONCEPTUAL UNDERSTANDINGS

1. The conditions of the demand side and the supply side of the economy are subject to continuous *change*, which causes variations in economic activity over time.
2. Fluctuations in economic activity have major effects on the ***economic well-being*** of societies and the individuals within them.
3. Different schools of thought offer different explanations of macroeconomic problems and suggest different solutions for dealing with these.

These topics are addressed in the next five chapters. Chapter 8 is concerned with the causes of changes in economic activity and the problem of how economists measure economic well-being. Chapter 9 introduces fundamental macroeconomic concepts that we will use to analyse the macroeconomy, in particular aggregate demand and the more controversial aggregate supply. Chapters 10 and 11 will discuss the important macro issues of inflation, unemployment, economic growth and debt. Finally, Chapter 12 deals with the challenges posed by increasing inequalities and how these may be effectively addressed.

Madrid, Spain. Job seekers outside an employment agency

> Chapter 8

The level of overall economic activity

BEFORE YOU START

- In the news you have probably heard reports describing the state of 'the economy'. What do you think 'the economy' is?
- News reports often comment on how the economy has improved or worsened. What do you think characterises an economy that has improved or worsened?

In this chapter we will discover how economists measure an economy's total output and income. We will study the business cycle, that shows how economic activity fluctuates over time. Finally, we will examine the limitations of standard measurements of output and income, and will consider alternative methods that may be more appropriate for measuring economic well-being.

Photo: Man pays for an order in a café with a credit card

8.1 Economic activity

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- review the circular flow of income model introduced in Chapter 1 and draw a diagram of this model (AO2, AO4)

The circular flow of income model

The circular flow of income model was introduced in Chapter 1, Section 1.3. We will briefly review it here as it forms the basis for understanding the macroeconomy.

The model in a closed economy with no government

Figure 8.1 is the same as Figure 1.4 (Chapter 1), which introduced the relationship and interdependence between households (consumers) and firms (businesses), linked together through product markets and resource markets.

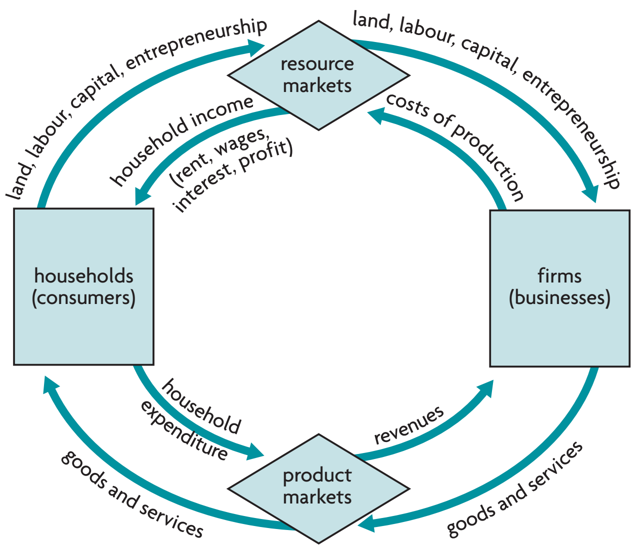


Figure 8.1: Circular flow of income model with no leakages and injections

Households, as owners of factors of production (land, labour, capital and entrepreneurship) sell these to firms, and they also buy the products that firms produce. Firms buy the factors of production, and sell the goods and services they produce to consumers. There is therefore a flow in the clockwise direction of factors of production from households to firms, and of goods and services from firms to households.

The counterclockwise direction shows flows of *money*. Households receive income when they sell their factors of production to firms in the form of *rent* (for land), *wages* (for labour), *interest* (for capital) and *profit* (for entrepreneurship). Consumers then have *household expenditures* which is the money they spend to buy goods and services. Firms on the other hand have *costs of production* when they buy the factors of production, and they receive *revenues* when they sell their goods and services.

We thus see that the *income flow* from firms to households is equal to the *expenditure flow* from households to firms: the household incomes from the sale of factors of production equals household expenditures on goods and services. This is the *circular flow of income*.

These two flows are also equal to the value of goods and services, or the *value of output flow*. This is equal to the sum of the product of each good and service multiplied by its respective price, giving the value of total output. Therefore:

> The circular flow of income shows that in any given time period (say a year), the *value of output* produced in an economy is equal to the *total income* generated in producing that output, which is equal to the *expenditures* made to purchase that output.

Adding leakages and injections

Figure 8.2 is the same as Figure 1.6, showing the circular flow of income model with *injections* (money flowing in) and *leakages* (money flowing out, also known as *withdrawals*) to the money flow of Figure 8.1.

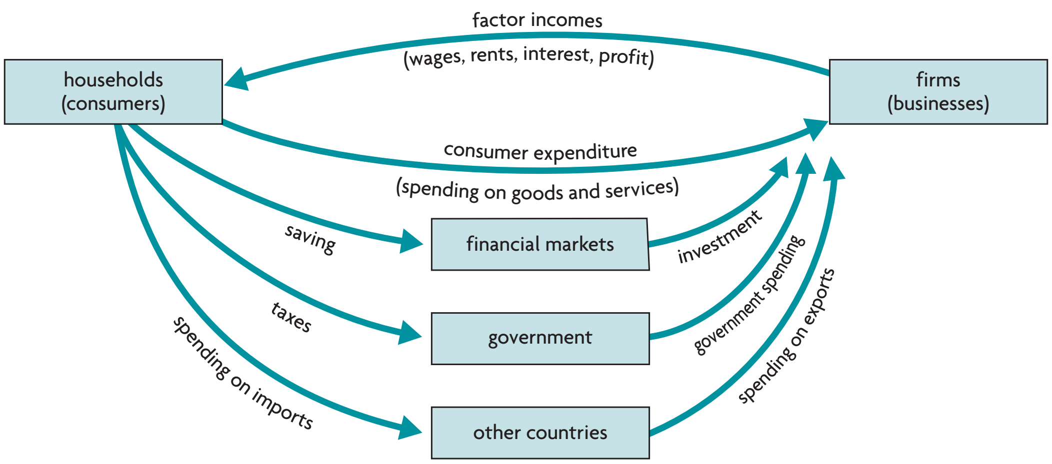


Figure 8.2: Circular flow of income model with leakages and injections

Leakages and injections are paired together so that what leaks out of the flow can come back in as an injection:

leakages	**injections**
saving	investment
taxes	government spending
import spending	export spending

Saving and investment

Saving is a part of consumer income that is not spent, therefore it is a leakage. Investment is spending mainly by firms for the production of capital goods (one of the four factors of production, physical capital). This is why capital goods are also known as investment goods. Households place their savings in financial markets (bank accounts, purchases of stocks and bonds, etc.) and firms obtain funds from financial markets (through borrowing, issuing stocks and bonds, etc.) to finance investment, or the production of capital goods. These funds therefore flow back into the expenditure flow as injections. Saving leaks out of the flow of consumer expenditures, passes through financial markets, and is then injected back into the expenditure flow as investment.

Taxes and government spending

Taxes and government spending are connected through the government. Taxes are a leakage since households pay taxes to the government instead of buying goods and services, and government spending on various activities (education, health, defence, etc.) comes back into the flow as an injection.

Imports and exports

An economy that imports and exports through international trade is an *open economy*. Imports and exports are linked together through 'other countries'. Imports represent a leakage because they are household spending that leaks out to the other countries that produce the goods and services. Exports represent an injection because they are spending by foreigners who buy domestically produced goods and services.

The size of the circular flow in relation to the size of leakages and injections

The relative size of leakages and injections has important consequences for the size of the circular flow. If leakages are greater than injections, the size of the circular flow becomes smaller. Suppose saving is larger than investment, so that the household income that leaks as saving into financial markets does not all come back into the flow as investment. The result is that fewer goods and services are purchased, firms cut back on their output, they buy fewer factors of production, unemployment increases (since firms buy a smaller quantity of labour) and household income is reduced.

If injections are larger than leakages, the size of the circular flow becomes larger. Suppose spending on exports is greater than spending on imports. This means that foreigners demand a greater amount of goods and services, firms begin to produce more by purchasing more factors of production, unemployment falls (as firms buy a larger quantity of labour), and household income increases.

To summarise:

Leakages from the circular flow of income (saving, taxes and imports) are matched by injections into the circular flow of income (investment, government spending and exports), though these need not be equal to each other. If injections are smaller than leakages, the income flow becomes smaller; if injections are larger than leakages the income flow becomes larger.

TEST YOUR UNDERSTANDING 8.1

It is suggested that you return to Test your understanding 1.10 in Chapter 1 to review the main points of the circular flow of income model.

8.2 Measures of economic activity

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the expenditure, income and output approaches to measuring GDP in national income accounting, and show their equivalence using the circular flow of income model (AO2, AO4)
- explain the following measures (AO2)
    - nominal Gross domestic product (GDP)
    - nominal Gross national income (GNI)
    - real GDP and real GNI
    - real GDP and real GNI per person (per capita)
    - real GDP and real GNI per person (per capita) at purchasing power parity (PPP)

Understanding measures of economic activity

Introduction to measures of economic activity

Measurement of economic activity involves measuring an economy's national income or the value of output, and is referred to as **national income accounting**. The output of an economy is referred to as **national output**, also known as *aggregate output*, which means total output. Knowing national income and the value of aggregate output is very useful because this allows us to:

- assess an economy's performance over time (are income and output increasing over time; are they decreasing?)
- make comparisons of income and output performance with other economies
- establish a basis for making policies that will meet economic objectives.

You may have noticed that we often refer to the 'value' of output. Why speak in terms of values, and not in terms of quantities, as we did in microeconomics? The answer is that in macroeconomics we must find a way to add up quantities of output of hundreds of thousands of different goods and services. Yet how can we add up quantities of computers, apples, cars and theatre tickets? What unit of measurement can we use? To get around this difficulty, we measure output in money terms, or the value of goods and services. The 'value' of a good is simply its quantity multiplied by its price. Sometimes 'value' may not be explicitly mentioned. For example, one may speak of 'the level of aggregate output' or simply 'aggregate output'. Whatever is the case, in macroeconomics output is always in value terms.

How economic activity is measured

The circular flow of income model showed that the value of aggregate output produced is equal to the total income generated in producing that output, which is equal to the expenditures made to purchase that output. For this reason, the term **national income**, or the total income of an economy, is sometimes used interchangeably with the value of *aggregate output*. We will now use this principle to see how national income or the value of aggregate output is measured.

There are three ways to measure the value of national output (or aggregate output), suggested by the circular flow of income model, all giving rise to the same result:

- the expenditure approach adds up all spending to buy final goods and services produced within a country over a time period
- the income approach adds up all income earned by the factors of production that produce all goods and services within a country over a time period
- the output approach calculates the value of all final goods and services produced in a country over a time period.

The expenditure approach

The expenditure approach measures the total amount of spending to buy final goods and services in a country (usually within a year). The term 'final' refers to goods and services ready for final use, and can be contrasted with intermediate goods and services, or those purchased as inputs for the production of final goods. When we measure the value of aggregate output, we include only purchases of final goods and services. For example, food items like meat and vegetables are intermediate goods for a restaurant that uses them to prepare a meal, and the meal is the final good. If in measuring expenditures we included spending on the food items plus spending on the meal, this would involve double counting and the value of aggregate output would be exaggerated. On the other hand, meat and vegetables bought by a household for consumption count as final goods, since they are not used as inputs for the production of another good or service.

Total spending is broken down into four components:

- **Consumption spending, abbreviated as *C*,** includes all purchases by households on final goods and services in a year (except housing, which is classified under investment).[1] Consumption spending is often referred to as **consumption** for short.
- **Investment spending, abbreviated as *I*,** includes:
  - spending by firms on capital goods (i.e. buildings, machinery, equipment, etc.)
  - spending on new construction (housing and other buildings).[2]

  Investment spending is often referred to as **investment** for short.
- **Government spending, abbreviated as *G*,** refers to spending by governments within a country (national, regional, local). It includes purchases by the government of factors of production, including labour services. It includes government investment, referred to as 'public investment', which usually involves roads, airports, power generators, buildings.
- **Net exports (exports minus imports), abbreviated as $X - M$,** refers to the value of all exports (abbreviated as X) minus the value of all imports (abbreviated as M). Exports are goods and services produced within the country and so must be included in the measurement of aggregate output. Imports, however, involve domestic spending on goods and services that have been produced in other countries, and so must be subtracted from expenditures measuring domestic output.

If we add together the four components of spending, we obtain a measure of aggregate output known as *gross domestic product* (*GDP*):

$$C + I + G + (X - M) = \text{GDP}$$

This approach allows economists to see the relative contribution of each component of GDP, how these might change over time, and to make comparisons over time or across countries. For example, if one is interested in investment, it is possible to trace out how the share of investment in GDP changes over time and to see how the share of investment in one country compares with its share in other countries.

[1] Spending by consumers is classified as spending on (i) consumer durable goods (with an expected life of more than three years, such as cars, refrigerators, washing machines, televisions, etc.), (ii) consumer non-durable goods (with an expected life of less than three years, such as food, clothing and medicines), and (iii) services (entertainment, banking, health care, education, etc.).

[2] Investment spending includes one more item: changes in inventories. Inventories refer to output produced by firms that remains unsold. Businesses as a rule keep inventories to help them meet unexpected increases in the demand for their product. Since inventories are output, it means that they must be counted as part of the aggregate output that is being measured. However, since they are output that has not been sold, they cannot be counted under consumption expenditure; they are therefore counted under investment. Investment spending is often referred to as **investment** for short.

Gross domestic product (GDP) is defined as the market value of all final goods and services produced in a country over a time period (usually a year). It includes spending by the four components, $C+I+G+(X-M)$. It is one of the most commonly used measures of the value of aggregate output.

Some clarifications concerning investment

Investment refers to spending by firms or the government on capital goods and on construction. There is a common misconception that investment is undertaken only by firms. This is incorrect, because as we have seen in the discussion above, investment is also undertaken by governments (public investments). One reason for this misconception is that firms are major decision-makers whose contribution to GDP is in the form of investment.

Another issue is that in measures of aggregate output, investment in 'capital' includes only spending on physical capital. It does not include spending on human capital and often it does not include spending on natural capital (see Chapter 1, Section 1.1 for these distinctions). This can be confusing, because economists often refer to 'investments in human capital' and 'investments in natural capital', and yet these 'investments' do not appear as investments in measures of aggregate output. Many economists argue that national income accounting methods should be changed to include these types of investments as well.

Note that measures of aggregate output also do not include financial capital or social capital (see Chapter 1, Section 1.1), but this is justifiable as these types of capital do not represent value of goods and services produced.

The income approach

The income approach adds up all income earned by the factors of production within a country over a time period (usually a year): wages earned by labour, rent earned by land, interest earned by capital and profits earned by entrepreneurship. When all factor incomes are added up, the result is *national income*. Whereas national income is often used as a measure of the level of economic activity, it is not the same as GDP. To calculate GDP using the income approach, it is necessary to make some adjustments to national income.[3]

This approach allows economists to see the relative income shares of the different factors of production, how these might change over time, and to make comparisons over time or across countries. For example, if we are interested in wages of workers, we can see how the share of wages in national income changes over time and how this share in one country compares with its share in other countries.

The output approach

The output approach measures the value of each good and service produced in the economy over a particular time period (usually a year) and then sums them up to obtain the total value of output produced. It includes the value of all final goods and services, in order to avoid the double counting that would arise from including the values of intermediate goods and services.[4]

The output approach calculates the value of output by economic sector, such as agriculture, manufacturing, transport, banking, etc. The value of output of each sector is then added up to obtain the total value of output for the entire economy.

This approach provides economists with the opportunity to study the performance of each individual sector

[3] A detailed consideration of these adjustments is beyond the scope of this book. For the interested student, they will be mentioned briefly here. If we add depreciation and indirect taxes to national income, we obtain a measure of aggregate output called gross national income (GNI). The difference between GNI and GDP will be considered later in this chapter. Depreciation refers to the wearing out of capital goods, and will also be considered later. The reason we add depreciation and indirect taxes to national income in order to obtain GNI (and GDP) is that the value of output measured by the expenditure approach includes both these items. By contrast, national income, measuring only the incomes of the factors of production, does not include either of the two.

[4] The method used to obtain the value of only final goods and services is to count only the value added in each step of the production process. For example, say the production of a good goes through the following steps. Firm A sells raw materials for $700 to firm B. Firm B uses the raw materials and produces an intermediate good that it sells to firm C for $1100. Firm C uses this intermediate good to produce a final good that it sells for $1700. How much value has been added in this process? Firm A added $700 of value. Firm B added $400 of value (= $1100 − $700), and firm C added $600 of value (= $1700 − $1100). When we add these up we obtain: $700 + $400 + $600 = $1700. Note that the sum of the values that were added in each step of the production process is exactly equal to the value of the final product. If we had added up the values of the two intermediate products and the final product, we would have: $700 + $1100 + $1700 = $3500, which greatly exaggerates the value of the product due to double counting. By counting only values added in each step of the production process, the problem of double counting is avoided.

by looking at its relative share in total output, how this changes over time and to make comparisons of performance across sectors across countries.

The three approaches give rise to the same result, after allowance is made for statistical differences that arise in the course of measuring the different variables involved.

TEST YOUR UNDERSTANDING 8.2

1. Explain why the terms 'national income' and 'aggregate output' are often used interchangeably.
2. Identify some reasons why it is useful to know the value of aggregate output.
3. Explain why
   a. we measure aggregate output in value terms, and
   b. we count only the value of final goods and services when measuring the value of output.
4. Identify the four expenditure components of GDP and explain each of these.
5. a. Explain three ways that GDP can be measured.
   b. Outline the type of information each approach offers about the economy.
6. Using the circular flow model, explain why the three ways to measure GDP give rise to the same result.

Distinctions relating to measures of the value of output

Distinction between GDP and GNI

Under the expenditure approach to measuring aggregate output, we learned the meaning of gross domestic product (GDP), which is the market value of all final goods and services produced in a country over a time period (usually a year), and includes the four components of spending: C + I + G + (X – M).

In addition, using the circular flow of income model, we learned that the value of output produced in an economy is equal to the total income generated in producing that output. However, in the real world, this equality does not always hold. Sometimes the output of an economy is produced by factors of production that belong to foreigners. Consider the case where a US multinational firm in India remits (sends back) its profits to the United States. The output of the multinational is produced in India, but the profit income is received by residents in the United States. Does the profit income count as Indian or US income and output? Consider also a Russian worker who lives and works in Spain, and sends a large part of her income to her family in Russia. Her output is produced in Spain, but the income she sends home is Russian income; should this income count as Russia's or Spain's income and output?

The concepts 'domestic' and 'national' are used to distinguish between measures of aggregate output and income that deal with this issue. The term 'domestic' in 'gross domestic product' means that output has been produced by factors of production domestically, or within the country, regardless of who owns them (residents or foreigners). The term 'national' is used in another measure of aggregate output known as **gross national income (GNI)**. The term 'national' in GNI means that the income it measures is the income of the country's residents, regardless where this income comes from.

In the first example above, the profit income remitted to the United States is included in Indian GDP because it is created by production in India, but it is part of US GNI because it is income received by US residents. The Russian worker's output in Spain is included in Spain's GDP, but her income sent to Russia is part of Russia's GNI.

GDP is the total value of all final goods and services produced within a country over a time period (usually a year), *regardless who owns the factors of production*. GNI is the total income received by the residents of a country, equal to the value of all final goods and services produced by the factors of production supplied by the country's residents *regardless where the factors are located*.

We will discuss GDP and GNI further in Chapter 18. It may interest you to turn to Table 18.2 to see some international comparisons.

Distinction between nominal values and real values

Earlier we noted that in macroeconomics we measure output in value terms, and we defined 'value' to be the quantity of a good multiplied by its price. **Nominal value** is money value, or value measured in terms of prices that prevail at the time of measurement. For example, if a pair of shoes costs £100, this is its nominal value. If you buy this pair of shoes, £100 is your nominal expenditure on these shoes. If your monthly income is £2000, this is your nominal income. Therefore, when we calculate the value of aggregate output, or expenditure, or income, in money terms, we speak of *nominal GDP*, *nominal GNI*, *nominal expenditure*, etc.

Yet prices change over time, and this poses a measurement problem. Let's say that nominal GDP increases in a year. This increase may be due to changes in the quantities of output produced, or changes in the prices of goods and services, or a combination of both. We do not know what part of the increase is due to changes in output and what part to changes in prices. Yet we are interested in knowing how much the *quantity* of goods and services has increased. We must therefore find a measure of GDP that is not influenced by price changes.

To eliminate the influence of changing prices on the value of output, we must calculate real values. **Real value** is a measure of value that takes into account changes in prices over time. Meaningful comparisons over time in the value of output, or expenditures, or income, or any variable that is measured in money terms, require the use of real values. For example, when we make comparisons of GDP in a country over time, we must be sure to use real GDP values, as these have eliminated the influence of price changes, and give us an indication of how actual output produced has changed.

> Nominal GDP and nominal GNI are measured in terms of current prices (prices at the time of measurement), which does not account for changes in prices. Real GDP and real GNI are measures of economic activity that have eliminated the influence of changes in prices. When a variable is being compared over time, it is important to use real values.

Distinction between total and *per capita* values

Per capita is a Latin expression that means per person. A *per capita* measure takes the total value (of output, income, expenditure, etc.) and divides this by the total population of a country. Therefore **GDP *per capita*** of a country is total GDP of that country divided by its population.

The distinction between total and *per capita* measures is very important for two reasons:

- **Differing population sizes across countries**. Let's say there are two countries that have identical total GDPs of £10 billion. Country A has a population of 1 million people and country B has a population of 2 million people. If we divide total GDP by population we get GDP *per capita* of £10 000 for country A and £5000 for country B. Whereas both countries have identical GDPs, country B's *per capita* GDP is only half that of country A, because of differing population sizes.
- **Population growth.** Changes in the size of GDP (or GNI) *per capita* over time depend very much on the relationship between growth in total GDP (or GNI) and growth in population. In general, if total GDP increases faster than the population, then GDP *per capita* increases. But if the country's population increases faster than total GDP, then GDP *per capita* falls.

> Total measures of the value of output and income (such as GDP and GNI), provide a summary statement of the overall size of an economy. *Per capita* figures are useful as a *summary measure* of the standard of living in a country, because they provide an indication of how much of total output or total income in the economy corresponds to each person in the population on average.

The meaning of real GDP/GNI *per capita* at purchasing power parity (PPP)

In order to be able to make accurate comparisons of real GDP/real GNI or real GDP *per capita*/real GNI *per capita across countries*, we need to make another distinction. The reason is that different countries have different *price levels*. This means that the same amount of money in a low-price country has greater purchasing power (can buy more things) than in a high-price country.

Suppose a candy bar costs €3 in a high-price country and the same candy bar costs €1.50 in a low-price country. This means that the €3 you have in your pocket can buy one candy bar in the high-price country and two candy bars in the low-price country. Your €3 has greater *purchasing power* in the low-price country. Purchasing power refers to the quantity of goods and services that can be bought with money. Clearly if we do not take price level differences into account, we will not get an accurate picture of differences across countries in the value of output they produce.

We therefore need a method of currency conversions (changing the currency of one country into the currency of another country) that accounts for different price levels, and therefore different purchasing powers across countries. Such a method is provided by special exchange rates called **purchasing power parities (PPPs)**. (You will learn about exchange rates in Chapter 16.) Purchasing power parity literally means 'buying power equivalence'. It is defined as the amount of a country's currency that is needed to buy the same quantity of local goods and services that can be bought with US$1 in the United States. The use of PPPs to make comparisons of any measure, whether GDP or GNI or anything else, eliminates the influence of price level differences and makes comparisons across countries far more accurate. (Use of the US$ as the basis for the conversions is only a matter of convention, as other currencies would also have been suitable for this purpose.)

Purchasing power parities will be discussed further in Chapter 18. You may be interested to see Table 18.3, which shows per capita GDP calculated both by use of standard exchange rates (column 1) and by use of purchasing power parities (column 2).

> Comparisons of GDP *per capita* (or GNI *per capita*) across countries require measures of *per capita* output or income based on conversions of national currencies into US$ by use of purchasing power parities (PPPs), to eliminate the influence of price differences on the value of output or income.

Purchasing power parity exchange rates are computed and published on a regular basis by several international bodies, including the Organisation for Economic Co-operation and Development (OECD), European Union, the World Bank and United Nations agencies

The meaning of *gross* in gross domestic product (Supplementary material)

If you are interested in learning about the meaning of *gross* in *gross domestic product* you will find the explanation in the digitial coursebook.

TEST YOUR UNDERSTANDING 8.3

1. a. Define GDP and GNI, and outline how they differ.
   b. Research and provide some examples of countries where (i) GNI is likely to be larger than GDP, and (ii) GDP is likely to be larger than GNI.
2. Explain the difference between nominal GDP and real GDP (or nominal and real GNI).
3. Outline why it is important to use real values when making comparisons over time.
4. You read in the newspaper that government spending on education in your country increased by 7% last year. Identify and describe further information that could help you make sense of this figure.
5. a. Outline the difference between total GDP (or GNI) and *per capita* GDP (or GNI).
   b. Explain why is it sometimes important to make a distinction between total measures and *per capita* measures of income and output.
6. a. Explain why price changes over time pose a problem when we want to make comparisons of GDP (or any measure of output or income) over time.
   b. Outline why the use of purchasing power parities (PPPs) is necessary to ensure that comparisons of GDP or GNI figures across countries are meaningful.

8.3 Calculations based on national income accounting

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text
- calculate the following measures (AO4)
  - nominal Gross domestic product (GDP) using the expenditure approach
  - nominal Gross national income (GNI)
  - real GDP and real GNI
  - real GDP and real GNI *per capita*

Calculating nominal GDP using the expenditure approach

We have seen that the measurement of GDP using the expenditure approach involves adding up the four spending components: consumption (C), firm investment (I), government spending (G) and exports minus imports ($X - M$). Therefore, $GDP = C + I + G + (X - M)$. (It is also possible to calculate GDP using the income and output approaches but this is beyond the scope of IB requirements.)

Suppose we are given the national income statistics for a country called Mountainland in Table 8.1 (Mnl is Mountainland's national currency).

Consumer spending	11.3
Investment spending	3.2
Government spending	3.5
Exports of goods and services	2.5
Imports of goods and services	2.1

Table 8.1: National income statistics for Mountainland, 2019 (billion Mnl)

Using this information, we find that nominal GDP $= 11.3 + 3.2 + 3.5 + 2.5 - 2.1 = 18.4$ billion Mnl in 2019. Note that all these figures are in nominal terms; therefore, this value of GDP is a *nominal* value.

Calculating GNI

The difference between GDP and GNI was explained earlier. Given data on GDP, we can find GNI in the following way: we add to GDP the income of domestic residents earned abroad, and subtract from GDP the income paid abroad to foreigners. Therefore:

GNI = GDP + income from abroad − income sent abroad

'Income from abroad – income sent abroad' can be simply written as 'net income from abroad'. We therefore have:

GNI = GDP + net income from abroad

Note that in the United Kingdom and in some other countries, 'net income from abroad' may be referred to as 'net property income from abroad'. In the United States, it is sometimes referred to as 'net foreign factor income' or 'net factor income'.

For example, suppose in 2020, Riverland's GDP was \$46 billion: income earned abroad and sent home to Riverland was \$2.7 billion; income earned in Riverland and sent abroad was \$4.7 billion. What was Riverland's 2020 GNI?

Riverland's net income from abroad was \$2.7 billion minus \$4.7 billion = –\$2 billion. Therefore,

2020 GNI = \$46 billion – \$2 billion = \$44 billion

Note that this value of GNI is a *nominal* value.

Calculating real GDP and real GNI using a price deflator

Understanding the difference between nominal and real GDP

The distinction between nominal and real values was discussed above. We will now use a numerical example to show how *real GDP* can be calculated from nominal GDP. This is normally done by statistical services in each country and our example here is for illustration purposes only (you will not have to perform such calculations).

Table 8.2 assumes a simple economy producing three items (burgers, haircuts and tractors). Part (a) shows their quantities and prices for three years and the corresponding nominal GDP. In 2001, 37 burgers selling at £3 each made the total value of burgers £111; 15 haircuts at £18 each had a value of £270; and 10 tractors at £50 each made the total value of tractors £500. Adding up the three total values, we find nominal GDP of £881 in 2001. The nominal GDP figures for 2002 and 2003 are calculated in the same way.

Part (b) of Table 8.2 shows that to find real GDP, it is only necessary to find the value of quantities produced in 2001, 2002 and 2003 using *the same prices of a single year*, called a *base year*. Any year can be used as the base year. In the table the base year is 2001. To calculate real GDP, we simply multiply the quantities of output produced each year by 2001 prices. Notice that columns 3, 6 and 9 are identical.

For example, in 2002, the 40 burgers are valued at the 2001 burger price of £3; the 17 haircuts are valued at the 2001 price of £18, and the 11 tractors are valued at the 2001 price of £50. Adding up the resulting values of the three items in column 7, we get a measure of real GDP of £976 in 2001 prices. Similarly, for 2003, the three quantities are also valued at the 2001 prices. Therefore, *real GDP is a measure of output valued at constant (unchanging) prices.*

Nominal GDP measures the value of current output valued at current prices, while real GDP measures the value of current output valued at constant (base year) prices.

a Calculating nominal GDP

1 Goods and services	2 2001 Q	3 2001 P	4 2001 value (Q × P)	5 2002 Q	6 2002 P	7 2002 value (Q × P)	8 2003 Q	9 2003 P	10 2003 value (Q × P)
Burgers	37	£3	£111	40	£4	£160	39	£5	£195
Haircuts	15	£18	£270	17	£20	£340	18	£21	£378
Tractors	10	£50	£500	11	£60	£660	10	£65	£650
Nominal GDP			£881			£1160			£1223

b Calculating real GDP

1 Goods and services	2 2001 Q	3 2001 P	4 2001 output in 2001 P (Q × P)	5 2002 Q	6 2001 P	7 2002 output in 2001 P (Q × P)	8 2003 Q	9 2001 P	10 2003 output in 2001 P (Q × P)
Burgers	37	£3	£111	40	£3	£120	39	£3	£117
Haircuts	15	£18	£270	17	£18	£306	18	£18	£324
Tractors	10	£50	£500	11	£50	£550	10	£50	£500
Real GDP			£881			£976			£941

Table 8.2: Nominal and real GDP in a hypothetical economy

Examining the changes in real GDP that occurred between 2001 and 2003, we find that real GDP increased from 2001 to 2002 (from £881 to £976), but decreased between 2002 and 2003, falling from £976 to £941. Note that real GDP fell in 2002–2003 even as nominal GDP increased over the same period; price increases caused nominal GDP to rise, while falling quantities meant that real GDP was falling.

Note also that in the base year, 2001, nominal GDP is equal to real GDP; this is always so for the base year since real GDP is valued at base year prices.

When we refer to real GDP figures, we must also refer to the specific base year used for the computation. In the example above, we say 'in 2003 real GDP at 2001 prices was £941'. The figure of £941 is otherwise meaningless, because if we had used a different base year, we would have arrived at a completely different figure for 2003 real GDP. It is also meaningless to compare real GDP figures calculated on the basis of different base years.

Understanding how the GDP deflator is derived

In the real world, the above method of converting nominal values into real values is extremely lengthy and complicated, as there are hundreds of thousands of products whose values must be measured. However, this is not a problem because economists use short-cut methods that take the form of price indices (*indices* is the plural of *index*). A *price index* is a measure of average prices in one period relative to average prices in a base year. A price index commonly used to convert nominal GDP to real GDP is a **price deflator** known as the *GDP deflator*:

$$GDP\ deflator = \frac{nominal\ GDP}{real\ GDP} \times 100$$

Statistical services derive the GDP deflator by using the values of nominal and real GDP they have already calculated (by the method in Table 8.2):

$$GDP\ deflator\ in\ 2001 = \frac{881}{881} \times 100 = 100.0$$

$$GDP\ deflator\ in\ 2002 = \frac{1160}{976} \times 100 = 118.8$$

$$GDP\ deflator\ in\ 2003 = \frac{1223}{941} \times 100 = 130.0$$

These results are summarised in Table 8.3. Note that the GDP deflator is 100.0 for 2001. *The index number for the base year is always equal to 100, for all indices*. This follows from the equality of nominal and real GDP in 2001, as we had selected 2001 to be the base year.

Year	Nominal GDP	Real GDP	GDP deflator
2001	£881	£881	100.0
2002	£1160	£976	118.8
2003	£1223	£941	130.0

Table 8.3: Nominal and real GDP

Using the GDP deflator to calculate real GDP

Statistical services in each country regularly publish GDP deflators (and other price indices). Using this information, it is a simple matter for economists to calculate real GDP from nominal GDP:

$$\text{real GDP} = \frac{\text{nominal GDP}}{\text{price deflator}} \times 100$$

For example, suppose we are given the following values of nominal GDP for a hypothetical Country X: \$7850 billion in 2001; \$9237 billion in 2002; and \$10 732 billion in 2003. We are also given the GDP deflator in Table 8.3, and are asked to calculate real GDP:

$$real\ GDP\ in\ 2001 = \frac{7850}{100.0} \times 100 = \$7850\ \text{billion}$$

$$real\ GDP\ in\ 2002 = \frac{9237}{118.8} \times 100 = \$7775\ \text{billion}$$

$$real\ GDP\ in\ 2003 = \frac{10732}{130.0} \times 100 = \$8255\ \text{billion}$$

Note that an increasing GDP deflator indicates rising prices on average, while a decreasing GDP deflator indicates falling prices on average. Suppose we have the following price index representing the GDP deflator:

2004	2005	2006	2007	2008
95.7	97.7	100.0	105.9	102.4

We can see that whereas prices on average increased in the period 2004–2007, in 2008 *they fell*. We can also see that the base year is 2006. Note that it is possible for some years to have a price index that is less than 100.0, which means simply that in those years, the average price level was lower than in the base year.

Calculating *per capita* values

Suppose that the population of Country X above was 310 million in 2001. We would like to calculate its real GDP *per capita* in that year:

$$\text{real GDP } \textit{per capita} = \frac{\$\,7850\,\textit{billion}}{310\,\textit{million}} = \$\,25\,323$$

Note that real GNI *per capita*, or any other measure *per capita* is calculated in exactly the same way.

TEST YOUR UNDERSTANDING 8.4

1 Calculate nominal GDP, given the following information from the national accounts of Flatland for the year 2019 (all figures are in billion Ftl, the national currency). Consumer spending = 125; government spending = 46; investment spending = 35; exports of goods and services = 12; imports of goods and services = 17.

2 Now suppose that profits of foreign multinational corporations in Flatland and incomes of foreign workers in Flatland that were sent home in 2019 were Ftl 3.7 billion. The profits of Flatland's multinational corporations abroad and income of Flatland workers abroad that were sent back to Flatland were Ftl 4.5 billion. What was Flatland's GNI in 2019?

3 You read in one source of information that real GDP in a hypothetical country in 2001 was $243 billion; in another source of information you read that real GDP in 2002 was $277 billion. State what information you need to be sure that the two figures can be compared with each other.

4 You are given the information in the table on an imaginary country called Lakeland.

Year	2015	2016	2017	2018	2019
Nominal GDP (billion Lkl)	19.9	20.7	21.9	22.6	22.3
Price deflator (GDP deflator)	98.5	100.0	102.3	107.6	103.7
Population million	1.20	1.21	1.22	1.23	1.27

a Identify the base year.

b Calculate real GDP for each of the five years in the table.

c State which year real GDP is the same as nominal GDP. Outline why.

d In 2017–2018, nominal GDP increased, but real GDP fell (check that this is what your calculations show). Explain how this could have happened.

e In 2018–2019, nominal GDP fell, but real GDP increased (check that this is what your calculations show). Explain how this could have happened.

f Calculate real GDP per capita for each of the years.

g In 2018–2019, real GDP increased but real GDP per capita fell (check that this is what your calculations show). Explain how this could have happened.

8.4 The business cycle

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the business cycle (AO2, AO4)
- distinguish between short-term fluctuations and the long-term growth trend (potential output) and draw a diagram illustrating these (AO2, AO4)

Understanding the business cycle

The cyclical pattern and phases of the business cycle

Whereas real output in most countries around the world grows over long periods of time, output growth virtually everywhere is uneven and irregular. In some years (or months) real output may grow rapidly, in other years (or months) more slowly, and in still others it may even fall, indicating negative growth.

A business cycle is shown in Figure 8.3, which plots real GDP on the vertical axis, against time on the horizontal axis. GDP is measured in real terms, so that the vertical axis measures changes in the volume of output produced after the influence of price-level changes has been eliminated. The cyclical line shows *actual output*, or real GDP that is actually achieved over time.

Business cycles consist of short-term fluctuations in the growth of real output, which are alternating periods of expansion (increasing real output) and contraction (decreasing real output).

Each cycle consists of the following phases:

- **Expansion.** An expansion occurs when there is positive growth in real GDP, shown by those parts of the curve in Figure 8.4 that slope upward. During expansions, employment of resources increases, and the general price level of the economy (which is an average over all prices) usually begins to rise more rapidly (this is known as inflation, to be discussed in Chapter 10).
- **Peak.** A peak represents the cycle's maximum real GDP, and marks the end of the expansion. When the economy reaches a peak, unemployment of resources has fallen substantially, and the general price level may be rising quite rapidly; the economy is likely to be experiencing inflation.
- **Contraction.** Following the peak, the economy begins to experience falling real GDP (negative growth), shown by the downward-sloping parts of the curve. If the contraction lasts six months (two quarters) or more, it is termed a **recession**, characterised by falling real GDP and growing unemployment of resources. Increases in the price level may slow down a lot, and it is even possible that prices in some sectors may begin to fall.
- **Trough**. A trough represents the cycle's minimum level of GDP, or the end of the contraction. There may now be widespread unemployment. A trough is followed by a new period of expansion (also known as a recovery), marking the beginning of a new cycle.

The term 'business cycle' suggests a phenomenon that is regular and predictable, whereas business cycles are in fact both irregular, as they do not occur at regular time intervals, and unpredictable. For these reasons, many economists prefer to call them 'short-term economic fluctuations'.

While each cycle typically lasts several years, it is not possible to generalise, as there is wide variation in how long the cycle lasts, as well as in intensity (how strong the expansion is and how deep the contraction or recession is). Expansions usually last longer than contractions. These are the reasons why the curve in Figure 8.3 has an irregular shape.

Short-term fluctuations and the long-term growth trend or potential output

The long-term growth trend and potential output

Figure 8.3 shows a line going through the cyclical line; this represents average growth over long periods of time (many years) and is known as the **long-term growth trend**. The long-term growth trend shows how output grows over time when cyclical fluctuations are ironed out. As you can see in the figure, real GDP actually achieved fluctuates around potential GDP (it fluctuates around the long-term growth trend).

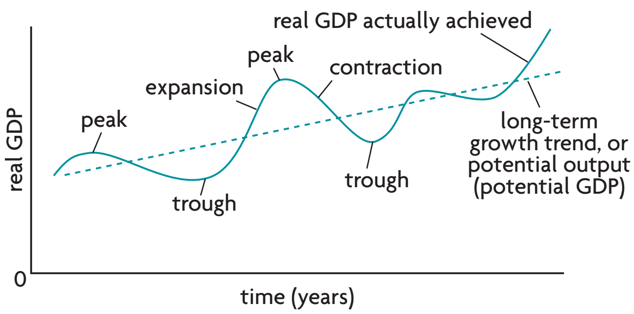


Figure 8.3: The business cycle

The output represented by the long-term growth trend is known as **potential output** or *potential GDP*. To understand the meaning of potential output, we must examine the relationship between real GDP and unemployment.

How unemployment relates to actual and potential output

When real GDP fluctuates, it does so together with other macroeconomic variables. One of the most important of these is unemployment of labour, or how many people in the workforce are out of work. When real GDP grows in the expansion phase, unemployment falls; in the contraction phase when real GDP falls, unemployment increases. You can easily see why: in an expansion, real GDP increases because firms produce more output; to do this, they hire more labour (and other resources) and unemployment falls. In a contraction, real GDP falls because firms cut back on production; as they lay off workers, unemployment increases.

For every economy, there is a level of real GDP at which the economy experiences 'full employment'. This is known as the full employment level of output, or full employment level of real GDP. The term **full employment** does not mean that all resources, including all labour resources, are employed to the greatest extent possible. Whenever the economy produces its 'full employment level of output', there is still some unemployment, known as the 'natural rate of unemployment'.

This is because at any time, there are some people who are in between jobs, some who are moving from one geographical area to another, some people who are training or retraining to be able to get a new or better job, and some people who are temporarily out of work. Therefore, there are always some people who are unemployed.

Coming back to potential output (shown by the long-term growth trend), we can now say that this is the level of output produced when there is 'full employment', meaning that unemployment is equal to the natural rate of unemployment. But when actual GDP is greater than potential GDP, unemployment is lower than the natural rate; when actual GDP is less than potential GDP, unemployment is greater than the natural rate.

Cyclical fluctuations, potential output and output gaps

Figure 8.4 introduces another concept related to the business cycle. When actual GDP lies above potential GDP (as at point d), or below potential GDP (as at point e), there results a *GDP gap*, also known as an *output gap*. The output gap is simply actual GDP minus potential GDP, and may be positive or negative. When actual GDP is equal to potential GDP (as at points a, b, c) the output gap is equal to zero.

The usefulness of these concepts will become apparent in later chapters when we make use of them to analyse short-term economic fluctuations and long-term growth.

Figure 8.4 shows that actual GDP fluctuates around full employment GDP, also known as potential GDP. When the economy's actual GDP is at points such as a, b and c, actual GDP is equal to potential GDP, and the economy is achieving full employment, where unemployment is equal to the **natural rate of unemployment**. When the economy's actual GDP is greater than potential GDP, such as at point d, there is an output gap, and unemployment falls to less than the natural rate. When actual GDP is less than potential GDP, such as at point e, there is an output gap where unemployment is greater than the natural rate.

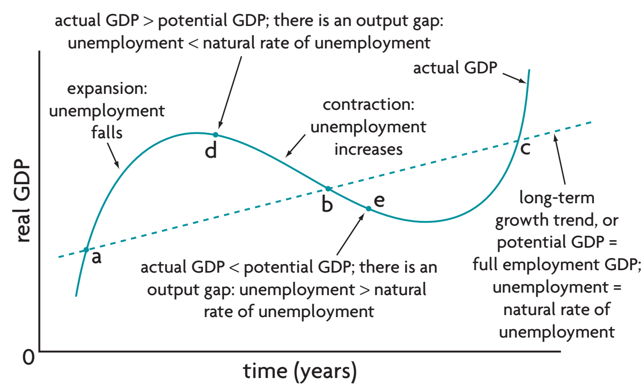


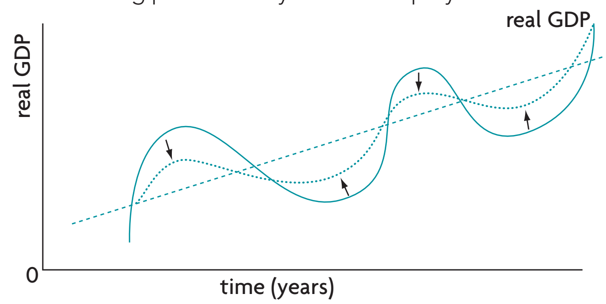
Figure 8.4: Illustrating actual output, potential output and unemployment in the business cycle

Why we study the business cycle

In an ideal world, every economy would experience economic growth over long periods of time, with continuous low levels of unemployment and a stable or gently rising price level (low inflation). Rapid economic growth, full employment, and price stability are among the key macroeconomic objectives of economies. Figure 8.5 illustrates these objectives in terms of the business cycle.

> Using the business cycle, we can understand macroeconomic objectives to include:
>
> - Reducing the intensity of expansions and contractions: this is aimed at making output gaps as small as possible (the dotted line in Figure 8.5(a)), by flattening the cyclical curve. This would lessen the problems of rising price levels or inflation in expansions and unemployment in contractions.
> - Increasing the steepness of the line representing potential output (the dotted line in Figure 8.5(b)), by achieving more rapid economic growth over long periods of time.

a Reducing the intensity of economic fluctuations: achieving price stability and full employment



b Economic growth

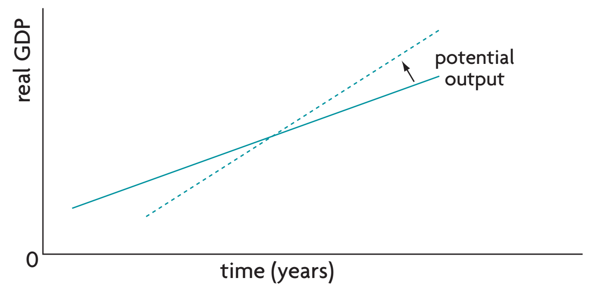


Figure 8.5: Illustrating three macroeconomic objectives

In the next chapter we will develop analytical tools to help us understand the causes of the business cycle, and in Chapter 13 we will study government policies intended to achieve full employment, price stability and economic growth.

TEST YOUR UNDERSTANDING 8.5

1. Using the business cycle diagram, distinguish between short-term fluctuations and the long-term growth trend.
2. Using a diagram:
   a. identify the phases of the business cycle,
   b. describe how they relate to unemployment, and
   c. distinguish between actual and potential output.
3. Describe how the 'natural rate of unemployment' relates to the 'full employment level of output' and to 'potential output'.
4. Suggest what an economy's business cycle is experiencing when there is
   a. a horizontal potential GDP line, and
   b. a downward-sloping potential GDP line.
5. Use a business cycle diagram to describe three important macroeconomic objectives.

THEORY OF KNOWLEDGE 8.1

The business cycle, actual output and potential output: using variables that cannot be observed

We have seen in our discussion of the business cycle (see Figure 8.4) that economists are concerned with two representations of output growth: growth of actual output and growth of potential output. Growth of actual output is straightforward to measure and show graphically; it consists of real GDP (or GNI) calculated for each year, and plotted on the vertical axis against time measured on the horizontal axis. When such data are plotted for any country over long periods of time, a cyclical pattern is likely to emerge (though an irregular one, for reasons explained in the text), known as the business cycle, or short-term fluctuations. Therefore, the empirical evidence supports the existence of a business cycle.

The case of potential output is different. Potential output is defined as full employment output, where unemployment is equal to the natural rate of unemployment, and its growth is shown by the long-term growth trend. However, for any particular economy, for any particular year, no one really knows what potential output is; nor does anyone know what the natural rate of unemployment is. Economists do not have any variable called 'potential output' or 'natural rate of unemployment' that they can observe in the real world and measure. Of course, economists make efforts to estimate the value of potential output (and the natural rate of unemployment), and to estimate how potential output changes over time. This then raises the question, does potential output actually *exist*, in the way that actual output can be said to exist, or is it a mythical idea that economists have created to help with their analysis of the macro economy?

It is not possible to provide a definite answer to this question; since potential output cannot be observed or measured, we cannot know if it exists. However, when theorising, there is nothing wrong with assuming the existence of something that cannot be directly observed; in other words, something whose existence is not supported by direct evidence. (Note that this is very different from making unrealistic assumptions, which conflict with the real world.) Physicists do this sometimes, with success. For example, to explain an event at the sub-atomic level (within atoms) it was necessary to presume the existence of a particle, though there was no direct evidence that such a particle actually existed. This fictional particle was named a neutrino, and 20 years later the neutrino was experimentally detected.

Sometimes, the unobserved variable may be supported by indirect evidence. For example, say we cannot observe X, but if X exists, then it is likely that Y exists; if we can observe and measure Y, then we can infer some characteristics about X. In the case of the neutrino, its existence was inferred from indirect evidence. (Note, however, that inference is not a full-proof method to arrive at conclusions about something, and may lead to wrong conclusions. For example, it may be true that if X exists, then Y also exists; but this does not necessarily mean that if Y exists, then X exists. To understand why, suppose that X = it is raining and Y = it is cloudy. If X is true (it is raining), then Y is true (it is cloudy). But if Y is true (it is cloudy), X (it is raining) is not necessarily true.)

Some economists argue that the existence of potential output is supported by indirect evidence, which may be helpful in making estimates about its size. Estimates of potential output can be useful to economists concerned with economic policy.

Thinking points

- Can you think of other variables used by economists that are not directly observable or measurable?
- Do you think the inability to observe some variables makes the social scientific method less 'scientific'?
- What kinds of difficulties might be created for policy-makers who use the concept of 'potential output' to determine appropriate policies for the economy?

8.5 National income statistics and alternative measures

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text
- evaluate the appropriateness of GDP or GNI statistics for the purpose of (AO3)
  - measuring economic well-being over time
  - comparing economic well-being across countries
- explain alternative measures of well-being including (AO2)
  - the OECD Better Life Index
  - the Happiness Index
  - the Happy Planet Index

Evaluating national income statistics

When real *per capita* GDP or real *per capita* GNI of a country increases over time, we might expect that the population of this country improves economic well-being. Alternatively, if GDP *per capita* or GNI *per capita* in one country is higher than in another country, we might expect that the first country enjoys a higher economic well-being. But would these conclusions be valid?

The answer is that we cannot be sure. There are two reasons why this is so. One is that **national income statistics** (or statistical data used to measure national income and output and other measures of economic performance) do not accurately measure the 'true' value of output produced in an economy. The other is that economic well-being is closely related to a variety of factors that GDP and GNI are unable to account for. As a result, *per capita* figures of both GDP and GNI may be misleading when used to make comparisons over time or comparisons between countries, and when used as the basis for economic well-being conclusions.

Why national income statistics (GDP/GNI) do not accurately measure the 'true' value of output

- **GDP and GNI do not include non-marketed output.** GDP measures the value of goods and services that are traded in the marketplace and that generate incomes for the factors of production. Yet some output of goods and services is not sold in the market and does not generate any income; this is called 'non-marketed output'. An example is one's own work on repairing and improving one's home; if the home repairs were carried out by hired workers, GDP would be greater by the amount of their wages. In developing countries households are often quite self-sufficient, with a substantial portion of production, such as agricultural production, taking place for a household's own use and consumption, and never reaching the marketplace. Non-marketed output therefore is likely to be far greater in developing countries compared to more developed ones. Many countries attempt to arrive at an estimate of non-marketed output, and by adding this to figures on marketed output arrive at a closer approximation of 'true' GDP/GNI.
- **GDP and GNI do not include output sold in underground (parallel) markets.** Here we have the case where goods are traded in markets and do generate incomes, but they go unrecorded and therefore are not included in GDP/GNI. An 'underground market' (also known as a 'parallel' or an 'informal market') exists whenever a buying/selling transaction is unrecorded. It may involve the sale of legal goods and services, such as reselling a good at a higher price if there is a price ceiling (see Chapter 4); or when a plumber does repairs in your home and does not report the income received to avoid paying taxes. Alternatively, it may include transactions involving illegal goods and services (such as drugs). In these cases as well, estimates of the size of underground markets can be made, and when added to the official (or recorded) economy, can arrive at a closer approximation of 'true' GDP and GNI.
- **GDP and GNI do not take into account quality improvements in goods and services.** The quality of many products improves over time, yet this is not taken into account in calculating the value of

total output. Technological advances often permit improved products to be sold at lower prices (for example, mobile phones and computers). This process offers significant benefits to consumers, which do not show up in GDP and GNI figures.

- **GDP and GNI do not account for the value of negative externalities, such as pollution, toxic wastes and other undesirable by-products of production.** Virtually all countries contribute to environmental degradation, reducing society's well-being, though this is not reflected in GDP/GNI figures.
- **GDP and GNI do not take into account the depletion of natural resources.** The depletion of natural resources (rainforests, wildlife, agricultural soils, etc.) also reduces society's well-being, yet is not taken into consideration.
- **GDP and GNI and differing domestic price levels.** Goods and services often sell for very different prices in different countries. If international comparisons of GDP do not account for differing price levels across countries, the result is a highly misleading picture of standards of living in different countries. This problem can be effectively dealt with if we convert values of GDP and GNI of different countries into a single common currency by use of *purchasing power parities* that take into consideration the differing price levels. Purchasing power parities were discussed above.

Why measures of the value of output (GDP/GNI) cannot accurately measure economic well-being

- **GDP and GNI make no distinctions about the composition of output.** Whether a country produces military goods (weapons, guns, tanks, etc.) or merit goods (education, health care, clean water supplies, and other services) or any other type of goods, GDP and GNI include the value of all without any distinctions about the degree to which they contribute to standards of living. One country may have a lower *per capita* GDP than another, but higher levels of social services and merit goods provision than the other. Which has higher standards of living? The GDP and GNI measures are unable to provide an indication.
- **GDP and GNI cannot reflect achievements in levels of education, health and life expectancy.** A society's levels of health and education contribute significantly to standards of living. Countries may achieve higher or lower levels of health and education with a given amount of GDP/GNI *per capita*, but these remain unaccounted for in measures of GDP and GNI. Increased life expectancy (the number of years one can expect to live, on average) is another benefit of technological improvements, improved health and higher income levels that has contributed enormously to a higher standard of living, but is not accounted for in GDP and GNI figures.
- **GDP and GNI provide no information on the distribution of income and output.** How equally or unequally income and output are distributed is another factor underlying society's well-being. Are the wealth and income of a nation highly concentrated in relatively few hands while large portions of the population are unable to satisfy their basic needs, or are these relatively more equally distributed? Are the benefits of a growing GDP concentrated among a small group of beneficiaries, or are they widely distributed? Are inequalities increasing or decreasing? Measures of GDP or GNI *per capita* cannot address any of these questions, as they only provide an indication of *average output* or *average income* per person.
- **GDP and GNI do not take into account increased leisure.** In many countries around the world the average number of hours worked per week has declined significantly, with the number of hours of leisure correspondingly increasing. This contributes to society's standard of living, yet is not accounted for in GDP or GNI.
- **GDP and GNI do not account for quality of life factors.** A society's well-being depends upon a number of non-economic factors, such as the crime rate, a sense of security and peace arising from relations with other countries, well-functioning institutions, stress levels from working conditions, insecurities arising from uncertainties relating to one's job, the degree of political freedom, and many others. GDP and GNI cannot account for any of these.

National income measures and comparisons of economic well-being over time and between countries

Comparisons over time

Earlier in the chapter we saw that to make comparisons over time, we must use *real values* of income and output measures, which take into account changes in the price level over time. Yet even when using real values of income and output, it is clear from the discussion above that comparisons of real GDP/GNI over time may be misleading. An increase in real GDP of some percentage for a particular country may overestimate or underestimate the true change in the population's economic well-being because of such factors as improved product quality, improvements in health and education, increased leisure, improvements (or deterioration) in quality of life factors, possible changes in the value of non-marketed output, or in the size of underground markets, and so on.

Comparisons between countries

Both the inability of GDP/GNI to measure the true value of output, and the exclusion of many factors that contribute to economic well-being, similarly contribute to limiting the validity of international comparisons by use of these measures. For example, one country may have a high level of GDP *per capita*, which is concentrated among a small percentage of the population, while another may have a lower level of GDP *per capita*, which is more equally distributed. A comparison of GDP/GNI figures will not reveal any information on this point, as well as on the other points listed above.

TEST YOUR UNDERSTANDING 8.6

1. Explain some reasons why national income statistics do not measure the 'true' value of income and output.
2. Explain some reasons why GDP *per capita* or GNI *per capita* may be inappropriate as the basis for making comparisons of a population's economic well-being over time, or comparisons between countries.

Alternative measures of well-being

In response to growing concerns that national income accounting measures such as GDP and GNI do not accurately reflect economic well-being, several alternative measures have been developed that try to capture more factors that affect well-being and quality of life.

OECD Better Life Index

The Organisation of Economic Co-operation and Development (OECD) is an intergovernmental organisation established in 1961, consisting of 36 member countries as of 2020. Most of the members are economically more developed countries. Its main purpose is to provide a forum for member countries to discuss common problems and policies and to promote policies that will encourage economic well-being.

The OECD has developed the **OECD Better Life Index** that is based on a number of factors that the member countries themselves selected as factors that make a better life. The purpose of this measure is to provide a more accurate representation of well-being and to form the basis of policies intended to improve the quality of life and well-being more generally. As the OECD notes

> *'Societal progress is about improvements in the well-being of people and households. Assessing such progress requires looking not only at the functioning of the economic system but also at the diverse experiences and living conditions of people.'*[5]

Figure 8.6 shows that according to the OECD there are two groups of factors that determine well-being in the present: quality of life which is measured in eight dimensions or indicators, and material conditions which are measured in three. Both quality of life factors and material conditions depend on four types of capital in the future. Natural capital refers to environmental resources; human capital refers to levels of education, skills and health; economic capital refers to money and wealth; and social capital refers to networks of people with shared values and understandings that facilitate co-operation. These four types of capital ensure that there will be sufficient resources in the future in order for a society to be able to maintain the well-being of its population.

Well-being in the present

Quality of life	Material conditions
State of health	Income and wealth
Work-life balance	Jobs and earnings
Education and skills	Housing
Social connections	
Civic engagement/governance	
Environmental quality	
Personal secuirty	
Subjective well-being	

Resources for well-being in the future

Natural capital	Human capital
Economic capital	Social capital

Figure 8.6: OECD framework for measuring well-being and progress

[5] http://www.oecd.org/statistics/measuring-well-being-and-progress.htm

Using the eleven indicators shown in Figure 8.6, the OECD constructs an index that ranks the countries according to their performance. Each year since 2011 it publishes the rankings of the countries and in addition shows how each country fares with respect to each of the eleven dimensions.

The OECD Better Life Index is still under development as different dimensions are sometimes added or taken out. For example it has been criticised for not taking equity factors into consideration. At the time of writing the OECD is working on how to incorporate equity into its Index. In addition it is likely that the Index suffers from difficulties in measurement of several of its dimensions, which may make the validity of comparisons across countries somewhat questionable.

Happiness Index

The **Happiness Index** began to be compiled in 2012 by the United Nations Sustainable Development Network. This is an organisation focused on gathering scientific and technological knowledge to encourage policies for sustainable development, including implementation of the Sustainable Development Goals (see Chapter 18) and the Paris Climate Agreement (see Chapter 5). The goal is to address the interdependent economic, social and environmental challenges faced by the world. In 2019 the Happiness Index included 156 countries.

The Happiness Index is based on the following dimensions:

- real GDP *per capita*
- social support
- healthy life expectancy
- freedom to make life choices
- generosity
- perceptions of corruption.

Data from all the participating countries are compiled from the Gallup World Poll which collects statistics based on telephone surveys in countries around the world on numerous topics like well-being, employment, access to food and many more. *The World Happiness Report* ranks countries according to the 'happiness' of their populations. In 2019, 156 countries were included. In addition, in 2018, countries were ranked by the happiness of their immigrants. The Report gives each country a rank from one to ten for each of the dimensions above, with ten being the best and one the worst. The countries are also given an overall rank summarising their performance in all the dimensions.

The results are published each year in *The World Happiness Report*. In addition to presenting the country rankings, the report each year focuses on a theme topic and its relation to happiness. In 2019, for example, it was happiness and community, in 2018 it was migration and happiness, and in 2017 it was the social foundations of happiness.

The Happiness Index has been criticised for limitations of some of the data and variables it uses, and in addition for being based on the concept of happiness. Happiness is very difficult to quantify and to measure. Happiness clearly means different things to different people, and its meaning varies across cultures, possibly making its rankings less reliable for comparisons across countries.

Happy Planet Index

The **Happy Planet Index** (HPI) was developed by the New Economics Foundation (NEF), a British non-governmental organisation (NGO, see Chapter 20) devoted to exploring new economic models 'based on equality, diversity and economic stability'.

In 2006, the NEF launched the Happy Planet Index (HPI) to challenge the idea that growth of GDP should be the most important goal of economic policy. It is argued that

> *'People vote for political parties that they perceive to be most capable of delivering a strong economy, and policy makers prioritise policies that increase in GDP as a result. Doing so has led to <u>short-termism, deteriorating social conditions, and paralysis</u> in the face of climate change.*
>
> *In fact, GDP growth on its own does not mean a better life for everyone, particularly in countries that are already wealthy. It does not reflect inequalities in material conditions between people in a country. It does not properly value the things that really matter to people like social relations, health, or how they spend their free time. And crucially, <u>ever-more economic growth is incompatible with the planetary limits we are up against. (emphasis in original)</u>.'*[6]

The Happy Planet Index is intended to be a measure of sustainable well-being. It takes into consideration life expectancy, how people feel about their own personal well-being, which are adjusted for inequalities and ecological footprint. It is calculated in the following way:

$$\text{Happy Planet Index (HPI)} = \frac{\text{life expectancy} \times \text{well-being} \times \text{inequality of outcomes}}{\text{ecological footprint}}$$

[6] http://happyplanetindex.org/about/#why

- *Life expectancy* is the average number of years a person expects to live, based on United Nations data.
- *Well-being* is taken to be a population's satisfaction measured by data collected by the Gallup World Poll
- *Inequality of outcomes* refers to inequalities between people with regard to life expectancy and well-being. Average well-being and life expectancy are adjusted downward to take into account inequalities in these dimensions.
- *Ecological footprint* is the impact on the environment of each individual in a society on average. It is measured as the amount of land needed to provide for all their requirements and the amount of land needed to absorb their CO_2 emissions. The higher the ecological footprint, the lower the HPI.

The HPI is calculated for 140–150 countries, depending on data availability. Each country receives a score from 0 to 100, the highest being the best.

Note that the *Happy Planet Index* and the *Happiness Index* refer to very different ideas. The Happiness Index is concerned with personal happiness while the Happy Planet Index is concerned with happiness of the planet. The Happy Planet Index is therefore much more of a measure of sustainability and how well resources can support a population's well-being.

The Happy Planet Index has been criticised for its measure of well-being, and it is also argued that the ecological footprint on which it is based is a controversial concept.

TEST YOUR UNDERSTANDING 8.7

1 Explain three alternative methods to measure well-being that have been put forward by various organisations.

2 In each of these cases, identify the factors that make them superior to standard national income statistics as a possible basis for comparisons over time or comparisons between countries.

THEORY OF KNOWLEDGE 8.2

The GDP concept

GDP as a concept was developed in the United States in 1934 (during the Great Depression) by US Nobel Prize winning economist Simon Kuznets.

After the Second World War, it became the main metric for measuring the size of a country's economy, its purpose being to measure the economy's ability to produce. However, since then, it has become a guide to policies to deal with numerous aspects of the economy, including inflation, unemployment, taxes, international trade and much more. In addition, it is used as an indicator of development, well-being and geopolitical strength. The US Commerce Department refers to it as 'one of the greatest inventions of the twentieth century'.

Yet Kuznets had warned about the limitations of the GDP concept:[7]

> *'The valuable capacity of the human mind to simplify a complex situation in a compact characterisation becomes dangerous when not controlled in terms of definitely stated criteria . . .*
> *Economic welfare cannot be adequately measured unless the personal distribution of income is known . . .*
> *The welfare of a nation can, therefore, scarcely be inferred from a measurement of national income as defined above.'*

[7] Simon Kuznets, 'Uses and Abuses of National Income Measurements', Report to the US Congress, 1934

CONTINUED

Many years later he wrote:[8]

> *'Distinctions must be kept in mind between quantity and quality of growth, between costs and returns, and between the short and long run. Goals for more growth should specify more growth of what and for what.'*

Much more recently Joseph Stiglitz, another US Nobel Prize winning economist wrote the following:[9]

> *'GDP is not a good measure of well-being. What we measure affects what we do: if we measure the wrong thing, we will do the wrong thing. If we focus only on material well-being – on, say, the production of goods, rather than on health, education, and the environment – we become distorted in the same way that these measures are distorted; we become more materialistic.'*

In fact GDP is a materialistic concept, according to which the overriding goal of any economy is greater production, without any regard as to whether it makes people better off. Yet in spite of its limitations, it is an attractive metric, because it is not political, and as a result, it allows governments and politicians to pursue it without having to resort to difficult questions about what *ought* to be society's goals. Everyone wants more output rather than less output, therefore pursuit of more output does not raise political or ideological objections. On the other hand, pursuit of greater income equality, environmental sustainability and more provision of merit goods including education and health care are politically highly controversial issues.

If society wants to pursue greater equality in income distribution, it must have a measure of well-being that includes measures of inequality. If society wants to pursue environmental sustainability, it must have a measure of well-being that incorporates achievements in this dimension. And so on with the numerous possible goals that a society might select.

But then, how is agreement, or political consensus to be reached within a society regarding which are the more important societal goals, since different groups may attach greater importance to some goals and less importance to others? Who is to decide on what particular dimensions should be included in a measure of well-being that will be used as the basis for policy?

Sources: *www.weforum.org; www.economist.com*

Thinking points

- Explain what Stiglitz means when he writes, 'What we measure affects what we do: if we measure the wrong thing, we will do the wrong thing.'
- As we know, economists tend to avoid normative issues in their thinking. To what extent do you think this is because value judgements should be kept separate from economic analysis? Do value judgements somehow 'contaminate' the impartial handling of economic facts and data?

[8] Simon Kuznets, 'How To Judge Quality'. The New Republic, 20 October 20 1962
[9] https://www.theguardian.com/business/2018/dec/03/gdp-wellbeing-health-education- environment-joseph-stiglitz

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1. Research GNI and GDP for various countries until you find a country with a large difference between the two values. Investigate to find what accounts for the difference.
2. Select five or more OECD countries you are interested in, research and find their real GDP or GNI per capita in $PPP, and rank them from highest to lowest. Find the corresponding countries in the OECD Better Life Index and rank these too from highest to lowest. Compare your two rankings in order to see the extent to which they differ. Examine each of the dimensions in order to suggest possible factors that might account for the differences between your GDP/GNI ranks and the OECD Better Life Index ranks.
3. Select five or more countries that appear in the Happiness Index and Happy Planet Index. Find their real GNI or GDP per capita in $PPP and rank them from highest to lowest. Find the corresponding countries in the Happiness Index and Happy Planet Index and rank these too from highest to lowest. You should now have three sets of ranks for your group of countries. Compare these three sets with each other in order to see the extent to which they differ. Suggest possible reasons why the rankings might differ from each other.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 9

Aggregate demand and aggregate supply

BEFORE YOU START

In the previous chapter you learned what 'growth' is and that there are short-term fluctuations and a long-term trend in growth.

1. What kinds of economic activities do you think cause short-term fluctuations in growth?
2. What kinds of economic activities do you think cause changes to the long-term growth trend?

In this chapter we will develop the aggregate demand–aggregate supply (*AD-AS*) model of the macroeconomy, an important analytical tool for studying output fluctuations, changes in the price level and unemployment, and economic growth.

Photo: Jinjiang, Quanzhou City, China. Workers at a factory making shoes

9.1 Aggregate demand (*AD*) and the aggregate demand curve

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the aggregate demand curve in terms of its components: consumption (*C*), investment (*I*), government spending (*G*), net exports (*X* – *M*) (AO2)
- explain the determinants of the components of aggregate demand: (AO2)
  - consumption (*C*): consumer confidence, interest rates, wealth, income taxes, level of household indebtedness, expectations of future price level
  - investment (*I*): interest rates, business confidence, technology, business taxes, level of corporate indebtedness
  - government (*G*): political and economic priorities
  - net exports (*X* – *M*): income of trading partners, exchange rates, trade policies
- explain shifts in the aggregate demand curve by reference to changes in the determinants of the components (AO2)
- draw the aggregate demand curve and shifts of the aggregate demand curve (AO4)

Explaining aggregate demand and the aggregate demand curve

The meaning of aggregate demand and the aggregate demand curve

Aggregate demand is the total quantity of aggregate (total) output, or real GDP, that all buyers in an economy want to buy at different possible price levels, *ceteris paribus*. The *aggregate demand* (*AD*) *curve* shows the relationship between the aggregate output buyers want to buy, or real GDP demanded, and the economy's price level, *ceteris paribus*. Figure 9.1(a) presents an aggregate demand curve. The horizontal axis measures aggregate output, or real GDP, and the vertical axis measures the general price level in the economy, which is an average over the prices of all goods and services.

Aggregate demand is not just the demand of all consumers, as one might think from the study of microeconomics. It consists of all the components of aggregate expenditure that we studied in Chapter 8, Section 8.2:

- the demand of consumers (*C*)
- the demand of businesses (firms) (*I*)
- the demand of government (*G*)
- the demand of foreigners for exports (*X*) minus the demand for imports (*M*) (*X* – *M* or net exports).

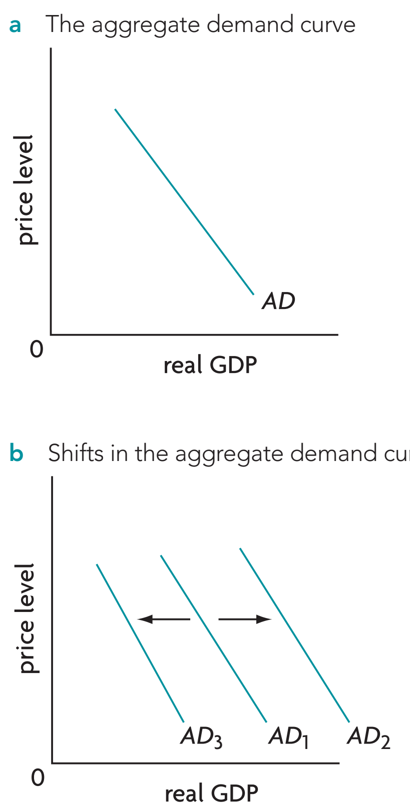


Figure 9.1: The aggregate demand (*AD*) curve

Aggregate demand is the total amount of real output (real GDP) that consumers, firms, the government and foreigners want to buy at each possible price level, over a particular time period. The **aggregate demand (*AD*) curve** shows the relationship between the total amount of real output demanded by the four components and the economy's price level over a particular time period. It is downward-sloping, indicating a negative relationship between the price level and aggregate output demanded.[1]

The negative (downward) slope of the aggregate demand curve (Supplementary material)

The reasons behind the downward slope of the aggregate demand are very different from demand in a single market in microeconomics. If you are interested in discovering the reasons behind the shape of the aggregate demand curve you may read about it in the digital coursebook.

The determinants of aggregate demand (shifts in the *AD* curve)

The meaning of aggregate demand curve shifts

It is important to distinguish between movements along the aggregate demand curve, caused by changes in the price level, discussed above, and shifts of the aggregate demand curve, caused by the **determinants of aggregate demand**, to which we turn next. (This is analogous to shifts of and movements along the demand curve in microeconomics.) Aggregate demand curve shifts are shown in Figure 9.1(b).

A rightward shift from AD_1 to AD_2 means that aggregate demand increases: for any price level, a larger amount of real GDP is demanded. A leftward shift from AD_1 to AD_3 means that aggregate demand decreases: for any price level, a smaller amount of real GDP is demanded.

Since aggregate demand is composed of consumer spending (C), investment spending (I), government spending (G), and net export spending ($X - M$), changes in aggregate demand and shifts in the aggregate demand curve can be caused by any factor that produces a change in one of these four components.

Causes of changes in consumption spending

- **Changes in consumer confidence.** Consumer confidence is a measure of how optimistic consumers are about their future income and the future of the economy. If consumers are optimistic about the future, they are likely to spend more on buying goods and services, and the *AD* curve shifts to the right. Low consumer confidence indicates expectations of falling incomes and worsening economic conditions, due to fears of cuts in wages or unemployment, causing decreases in spending, appearing as a leftward shift of the *AD* curve. Governments around the world regularly measure consumer confidence (through surveys based on questionnaires of consumers) to try to predict the level of consumer spending.
- **Changes in interest rates.** Some consumer spending is financed by borrowing, and so is influenced by interest rate changes. An increase in interest rates makes borrowing more expensive, resulting in lower consumer spending, and therefore a leftward shift in the *AD* curve. A fall in interest rates makes borrowing less expensive, and results in more consumer spending and a rightward shift in the *AD* curve. Interest rates can change as a result of a type of policy called 'monetary policy' (see Chapter 13).

[1] You may have noticed something odd about the definition of aggregate demand. In Chapter 8, Section 8.2 we defined GDP to be equal to spending by the four components: $C + I + G + (X - M)$. Now we are saying that aggregate demand is also equal to $C + I + G + (X - M)$. Yet aggregate demand is not the same as GDP. The explanation for this apparent oddity can be found in the digital coursebook 'Understanding aggregate demand and the multiplier in terms of the Keynesian cross model', included as Supplementary material.

- **Changes in wealth. Wealth** is the value of assets that people own, such as savings in their bank accounts, houses, stocks and bonds, jewellery, works of art, and so on; minus debt to banks or other financial institutions. An increase in consumer wealth (for example, an increase in the value of homes) makes people feel wealthier; therefore, they spend more and the *AD* curve shifts to the right. A decrease in wealth lowers aggregate demand; the *AD* curve shifts to the left.
- **Changes in income taxes.** If the government increases **income taxes** (taxes paid by households on their incomes), then consumer *disposable income*, which is the income left over after personal income taxes have been paid, falls; therefore, spending drops, and the *AD* curve shifts to the left. If personal income taxes are lowered, the result is higher disposable income and a rightward shift in the *AD* curve. Changes in taxes are the result of a type of government policy called 'fiscal policy' (see Chapter 13).
- **Changes in the level of household indebtedness.** 'Indebtedness' refers to how much money people owe from borrowing in the past. If consumers have a high level of debt (such as credit card use or taking out loans), then they are under pressure to make high monthly payments to pay back their loans plus interest, and so are likely to cut back on their present expenditures. Therefore, a high level of indebtedness lowers consumption spending and shifts the *AD* curve to the left. A low level of indebtedness increases consumption spending and shifts the *AD* curve to the right.
- **Expectations of future price levels**. Consumer spending may be influenced by what they expect prices to be in the future. If they expect prices of goods and services to fall, they may postpone spending as they wait for prices to fall, causing *AD* to decrease, shifting *AD* to the left. On the other hand if they expect future prices to increase, they may buy more now in order to avoid the higher prices later, thus causing *AD* to increase shifting to the right.

Causes of changes in investment spending

- **Changes in business confidence.** Business confidence refers to how optimistic firms are about their future sales and economic activity. If businesses are optimistic, they spend more on investment, and the *AD* curve shifts to the right. Business pessimism, on the other hand, results in a leftward shift in the *AD* curve.
- **Changes in interest rates.** Increases in interest rates raise the cost of borrowing, and force businesses to reduce investment spending financed by borrowing, and therefore the *AD* curve shifts to the left. Decreases in interest rates mean businesses can now finance their investment spending by borrowing at a lower cost, and the *AD* curve shifts to the right. As noted above, interest rates change as a result of monetary policy (see Chapter 13).
- **Changes (improvements) in technology.** Improvements in technology stimulate investment spending, thus causing increases in aggregate demand and a rightward shift in the *AD* curve.
- **Changes in business taxes.** Business taxes in this context refer to taxes on profits (also known as corporate income taxes). If the government increases taxes on profits of businesses (as part of its fiscal policy; see Chapter 13), firms' after-tax profits fall; therefore, investment spending decreases and the *AD* curve shifts to the left. Decreases in taxes on profits result in increased aggregate demand and a rightward *AD* curve shift.
- **The level of corporate indebtedness.** As in the case of household indebtedness, if businesses have high levels of debt due to past borrowing, they will be less willing to make investments and the *AD* curve shifts to the left. A low level of corporate indebtedness, on the other hand, leads to more investment and a rightward shift in the *AD* curve.
- **Legal/institutional changes.** Sometimes, the legal and institutional environment in which businesses operate has an impact on investment spending. This is often the case in many developing economies where laws and institutions do not favour small businesses. For example, small businesses often do not have access to credit, meaning they cannot borrow easily to finance investments. Many developing economies do not have the necessary laws that secure property rights (legal rights to ownership). In such situations, increasing access to credit (the ability to borrow) and securing property rights would result in increases in investment spending, shifting the *AD* curve to the right.

Causes of changes in government spending

- **Changes in political priorities.** Governments have many expenditures, arising from provision of merit goods and public goods, spending on subsidies and pensions, payments of wages and salaries to its employees, purchases of goods for its own use, and so on. It may decide to increase or decrease its expenditures in response to changes in its priorities. Increased government spending shifts the *AD* curve to the right, and decreased government spending shifts it to the left.

- **Changes in economic priorities: deliberate efforts to influence aggregate demand.** The government can use its own spending as part of a deliberate attempt to influence aggregate demand. The effects of such changes in government spending on aggregate demand are exactly the same as above. This is another aspect of fiscal policy (to be discussed in Chapter 13).

Causes of changes in export spending minus import spending

- **Changes in national income abroad.** Consider aggregate demand in country A, which has trade links with country B. If country B's national income increases, it will import more goods and services from country A, so that country A's exports will increase. Therefore the *AD* curve in country A shifts to the right. If, on the other hand, country B's national income falls, it will buy less from country A, and country A's *AD* curve shifts to the left.
- **Changes in exchange rates.** An exchange rate is the price of one country's currency in terms of another country's currency (see Chapter 16). Consider again country A, and assume that the price of its currency increases, becoming more expensive relative to the currency of country B. Country B now finds country A's output more expensive, and so it imports less from country A; therefore, country A's exports fall, and its *AD* curve shifts to the left. At the same time, country A now finds country B's output cheaper, and so it increases its imports from country B. Therefore, the increase in price of country A's currency has two effects: a fall in its exports and an increase in imports so that net exports, $X - M$, fall, and the *AD* curve shifts to the left. In the opposite situation, where the price of country A's currency decreases, an increase in exports and a decrease in imports will result, so that $X - M$ increases, and country A's *AD* curve shifts to the right.
- **Changes in trade policies, or the level of trade protection.** 'Trade protection' refers to restrictions to free international trade often imposed by governments (see Chapter 14). Suppose country A trades freely with country B (with no trade restrictions). However, country B's government decides to impose restrictions on imports from country A. Country A's exports will fall, and its *AD* curve will shift to the left. On the other hand, in country B, lower imports mean that the value of $X - M$ increases, and its *AD* curve shifts to the right.

Table 9.1 summarises the factors that can cause shifts of the aggregate demand curve.

Shifts in the aggregate demand curve are caused by:
Changes in consumer spending, arising from: • changes in consumer confidence • changes in interest rates (monetary policy) • changes in wealth • changes in personal income taxes (fiscal policy) • changes in the level of household indebtedness • expectations of future price levels
Changes in investment spending, arising from: • changes in business confidence • changes in interest rates (monetary policy) • changes (improvement) in technology • changes in business taxes (fiscal policy) • changes in the level of corporate indebtedness • legal/institutional changes
Changes in government spending, arising from: • changes in political priorities • changes in economic priorities: deliberate efforts to influence aggregate demand (fiscal policy)
Changes in foreigners' spending, arising from: • changes in national income abroad • changes in exchange rates • changes in the level of trade protection

Table 9.1: Factors causing shifts of the aggregate demand curve

TEST YOUR UNDERSTANDING 9.1

1 a Define aggregate demand and explain each of its four components.

b Show aggregate demand diagrammatically and define the relationship it represents.

2 Using diagrams, distinguish between a movement along the *AD* curve and a shift of the *AD* curve, and provide examples of the causes of each. Identify the four components of spending that cause shifts of the aggregate demand curve.

3 Using diagrams, show the impact of each of the following on the aggregate demand curve; explain what happens to aggregate demand in each case; and identify the component(s) of aggregate expenditure involved.

a Consumer confidence improves as consumers become optimistic about future economic conditions.

b The government decides to increase taxes on firms' profits.

c Firms become fearful that a recession is about to begin.

d The government decides to increase its spending on health care services.

e There is a decline in the real estate market (average house prices fall).

f The central bank (a government organisation) decides to increase interest rates.

g There is an increase in the level of indebtedness of consumers and firms.

h Real incomes in countries that purchase a large share of country A's exports fall; examine the impact on aggregate demand in country A.

i The government lowers personal income taxes (taxes on income of households).

j New legislation makes property rights more secure.

k There is an appreciation (an increase) in the value of the euro relative to the US dollar; examine the impact on aggregate demand in euro zone countries (countries that use the euro).

l There is an appreciation (an increase) in the value of the euro relative to the US dollar; examine the impact on aggregate demand in the United States.

m A non-governmental organisation (NGO) introduces a programme providing credit to small farmers, making it easier for small farmers to borrow to finance the building of irrigation projects.

Shifts in the *AD* curve and national income

Note that income is not included among the factors that can shift the *AD* curve. The reason is that *changes in national income cannot initiate any AD curve shifts*. This follows from the point noted earlier that real GDP, measured on the horizontal axis, also represents national income. It is not possible for any variable measured on either of the two axes to cause a shift of a curve (for an explanation, see the 'Quantitative techniques' chapter in the digital coursebook).[2] (This point will become clearer when we discuss the Keynesian multiplier at HL; see Chapter 13).

9.2 Short-run aggregate supply and short-run equilibrium in the *AD-AS* model

Note to the reader: The order of some topics in this section of the IB syllabus has been changed in order to facilitate presentation of the material. Specifically, short-run equilibrium is presented first, before consideration of 'Alternative views of aggregate supply (AS)'.

[2] Note that this does not contradict the ability of changes in disposable (or after-tax) income due to changes in taxes to affect aggregate demand. This is because changes in taxes and disposable income do not affect national income, as they simply involve a transfer of income from households to the government. National income remains unchanged.

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the short run aggregate supply (*SRAS*) curve (AO2)
- explain the determinants of the *SRAS* curve: (AO2)
    - costs of factors of production
    - indirect taxes
- explain shifts of the *SRAS* curve by reference to changes in the determinants (AO2)
- draw the *SRAS* curve and shifts of the *SRAS* curve (AO4)
- explain macroeconomic equilibrium in the short run (AO2)
- draw a diagram illustrating short-run equilibrium and changes in short-run equilibrium (AO4)

Whereas aggregate demand and the aggregate demand curve are straightforward and uncontroversial, aggregate supply is hotly debated by economists. Most of the disagreements focus on the shape of the aggregate supply curve. We will study three aggregate supply curves.

Short-run aggregate supply

The short run and long run in macroeconomics

The short run and the long run in macroeconomics differ from the corresponding distinction in microeconomics. The *short run in macroeconomics* is the period of time when prices of resources are roughly constant or inflexible, in spite of changes in the price level; they do not change together with changes in the price level. This applies especially to wages, or the price of labour. The *long run in macroeconomics* is the period of time when the prices of all resources, including the price of labour (wages), are flexible and *change along with changes in the price level*.

Wages, or the price of labour resources, are of special interest because they account for the largest part of firms' costs of production, and therefore strongly affect the quantity of output supplied by firms. Wages do not change very much over relatively short periods of time. The price of labour (wages) is often rigid (unchanging), because:

- labour contracts fix wage rates for certain periods of time, perhaps a year or two or more
- minimum wage legislation fixes the lowest legally permissible wage
- workers and labour unions resist wage cuts
- wage cuts have negative effects on worker morale, causing firms to avoid them.

The distinction between the short run and the long run in macroeconomics does not affect aggregate demand, but is very important for aggregate supply.

Defining aggregate supply and the short-run aggregate supply curve

We begin by defining aggregate supply and the short-run aggregate supply curve.

Aggregate supply is the total quantity of goods and services produced in an economy (real GDP) over a particular time period at different price levels.

The **short-run aggregate supply curve** (*SRAS*) shows the relationship between the price level and the quantity of real output (real GDP) produced by firms when resource prices (especially wages) do not change.

Figure 9.2(a) illustrates a short-run aggregate supply curve, indicating that there is a positive (or direct) relationship between the price level and real GDP supplied: a higher price level is associated with a greater quantity of real GDP, and a lower price level with a lower quantity of real GDP.

a The upward-sloping *SRAS* curve

b Shifts in the *SRAS* curve

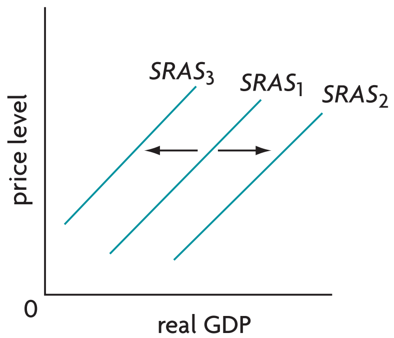


Figure 9.2: The short-run aggregate supply curve (*SRAS*)

Why the *SRAS* curve is upward-sloping

The positive relationship between the price level and real output (real GDP) relates to firm profitability: when there is an increase in the price level, this means that output prices have increased; but with unchanging resource prices (since the economy is in the short run), firms' profits increase. As production becomes more profitable, firms increase the quantity of output produced, resulting in the positive relationship between the price level and the quantity of real GDP supplied.

Similarly, a falling price level means falling output prices; with constant resource prices, firm profitability falls, and output decreases.

Changes in short-run aggregate supply (shifts in the *SRAS* curve)

A number of factors (other than the price level) cause shifts of the *SRAS* curve, illustrated in Figure 9.2(b).

This distinction is analogous to what we learned in Chapter 2 in connection with shifts of and movements along the supply curve for a specific good.

A rightward shift from $SRAS_1$ to $SRAS_2$ means that short-run aggregate supply increases: for any particular price level, firms produce a larger quantity of real GDP. A leftward shift from $SRAS_1$ to $SRAS_3$ means that aggregate supply decreases: for any particular price level, firms produce a smaller quantity of real GDP.

Important factors that cause *SRAS* curve shifts include:

- **Changes in wages.** If wages increase, with the price level constant, firms' costs of production rise, resulting in a leftward shift in the *SRAS* curve, such as from $SRAS_1$ to $SRAS_3$ in Figure 9.2(b). If wages decrease, with the price level constant, firms' costs drop, giving rise to a rightward shift in the *SRAS* curve, such as from $SRAS_1$ to $SRAS_2$.
- **Changes in non-labour resource prices.** Changes in the price of non-labour resources, such as the price of oil, equipment, capital goods, land inputs, and so on affect the *SRAS* curve in the same way as changes in wages. An increase in the price of a resource shifts the *SRAS* curve to the left; a decrease shifts it to the right.
- **Changes in indirect taxes.** Indirect taxes were studied in Chapters 4 and 5. They are treated by firms like costs of production. Therefore, higher indirect taxes are like increases in production costs and so shift the *SRAS* curve to the left. Lower indirect taxes on profits are like lower production costs and shift the *SRAS* curve to the right.[3]
- **Changes in subsidies offered to businesses.** Subsidies have the opposite effect to taxes, as they involve money transferred from the government to firms. If they increase, the *SRAS* curve shifts to the right; if they decrease, the *SRAS* curve shifts to the left.
- **Supply shocks.** Supply shocks are events that have a sudden and strong impact on short-run aggregate supply (see also Chapter 2, Section 2.3). For example, a war or violent conflict can result in destruction of physical capital and disruption of the economy, leading to lower output produced and

[3] You may remember that in Chapters 4 and 5 when we studied taxes and subsidies we shifted the supply curve upward and downward. A downward shift is equivalent to a rightward shift, and an upward shift is equivalent to a leftward shift. These same relationships apply to *aggregate* supply as well. See 'Quantitative techniques' in the digital coursebook for an explanation.

a leftward shift in the *SRAS* curve. Unfavourable weather conditions can cause a fall in agricultural output, also shifting the *SRAS* curve to the left. Beneficial supply shocks such as unusually good weather conditions with a positive effect on agricultural output lead to an increase in aggregate supply and a rightward shift in the *SRAS* curve.

Over short periods of time, the *SRAS* curve shifts to the left or to the right mainly as a result of factors that influence firms' costs of production (such as changes in wages, changes in non-labour resource prices and changes in business taxes or subsidies), as well as supply shocks.

TEST YOUR UNDERSTANDING 9.2

1 a Define aggregate supply.
  b Explain why the short-run aggregate supply curve is upward-sloping.

2 a Distinguish between the short run and the long run in macroeconomics.
  b What are some of the factors that cause wages to be inflexible (not change very easily and rapidly)?

3 a Show the short-run aggregate supply (*SRAS*) curve in a diagram, and explain what relationship it represents.
  b Identify the factors that can cause a movement along the *SRAS* curve.
  c Identify the factors that cause shifts in the *SRAS* curve.

4 Using diagrams, show the impact of each of the following on the *SRAS* curve; explain what happens to *SRAS* in each case.
  a The price of oil (an important input in production) increases.
  b Below-zero temperatures destroy agricultural output.
  c The government lowers taxes on firms' profits.
  d The government eliminates subsidies on agricultural products.
  e There is an increase in the minimum wage.

Short-run equilibrium in the *AD-AS* model

Illustrating short-run equilibrium

We will now put the aggregate demand curve and the short-run aggregate supply curve together, to determine short-run macroeconomic equilibrium.

In the *AD-AS* model, the **equilibrium level of output** occurs where aggregate demand intersects aggregate supply. **Short-run equilibrium** is given by the point of intersection of the *AD* and *SRAS* curves, and determines the price level, the level of real GDP and the level of employment.

This is shown in Figure 9.3 where Pl_e is the equilibrium price level and Y_e is the equilibrium level of real GDP.

At any price level and real GDP other than Pl_e and Y_e, the economy is in disequilibrium. At price level Pl_1, there is an excess amount of real GDP supplied, putting a downward pressure on the price level, which falls until it reaches Pl_e. At a price level lower than Pl_e, such as Pl_2, there is an excess amount of real GDP demanded, putting an upward pressure on the price level, which moves upward until it settles at Pl_e. At Pl_e, the amount of real GDP demanded is equal to the amount supplied, and there is short-run equilibrium.

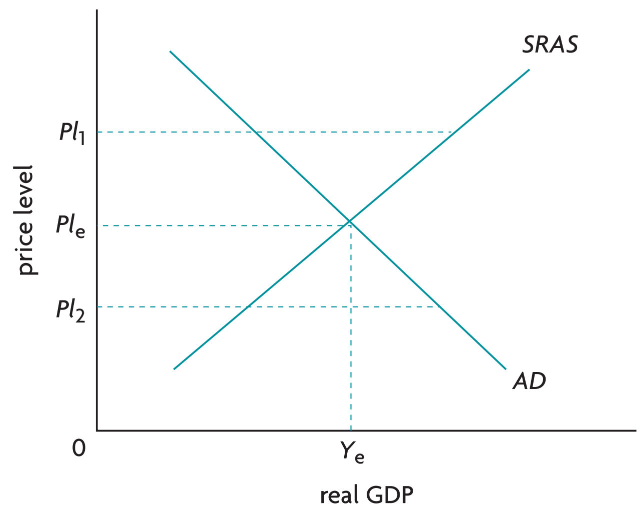


Figure 9.3: Short-run macroeconomic equilibrium

Impacts of changes in short-run equilibrium

The short-run equilibrium of an economy changes whenever there is a change in aggregate demand or short-run aggregate supply. Suppose there occurs an increase in aggregate demand, due to any of the factors discussed earlier (such as an increase in national income abroad, or an increase in investment spending). In Figure 9.4(a), the AD curve shifts from AD_1 to AD_2, resulting in an increase in the price level from Pl_1 to Pl_2, and an increase in real GDP, from Y_1 to Y_2. These changes also lead to a fall in unemployment since firms hire more labour in order to produce more output.

If there is a decrease in aggregate demand (due, for example, to pessimism among firms or a fall in net exports), the AD curve shifts from AD_1 to AD_3; the price level and real GDP fall from Pl_1 to Pl_3 and from Y_1 to Y_3, while unemployment increases since firms now need less labour.

Figure 9.4(b) shows shifts in the $SRAS$ curve. A rightward shift from $SRAS_1$ to $SRAS_2$ (for example, because of a technological improvement or lower business taxes), result in a lower price level, Pl_2, a higher level of real GDP, Y_2, and lower unemployment. On the other hand, a leftward shift from $SRAS_1$ to $SRAS_3$ (say, because of an increase in business taxes or an increase in the price of a resource) produce an increase in the price level to Pl_3, a fall in real output to Y_3 and an increase in unemployment.

TEST YOUR UNDERSTANDING 9.3

1 Using diagrams, show the effects of each of the following on short-run equilibrium, explaining what happens to the equilibrium price level, output and unemployment.

- a The price of oil (an important input in production) increases.
- b Firms are pessimistic about the future of the economy.
- c Below-zero temperatures destroy agricultural output.
- d The government lowers taxes on firms' profits.
- e There is a large rise in stock market prices.
- f The government eliminates subsidies on agricultural products.
- g A war destroys a portion of an economy's physical capital.
- h Consumer confidence improves.

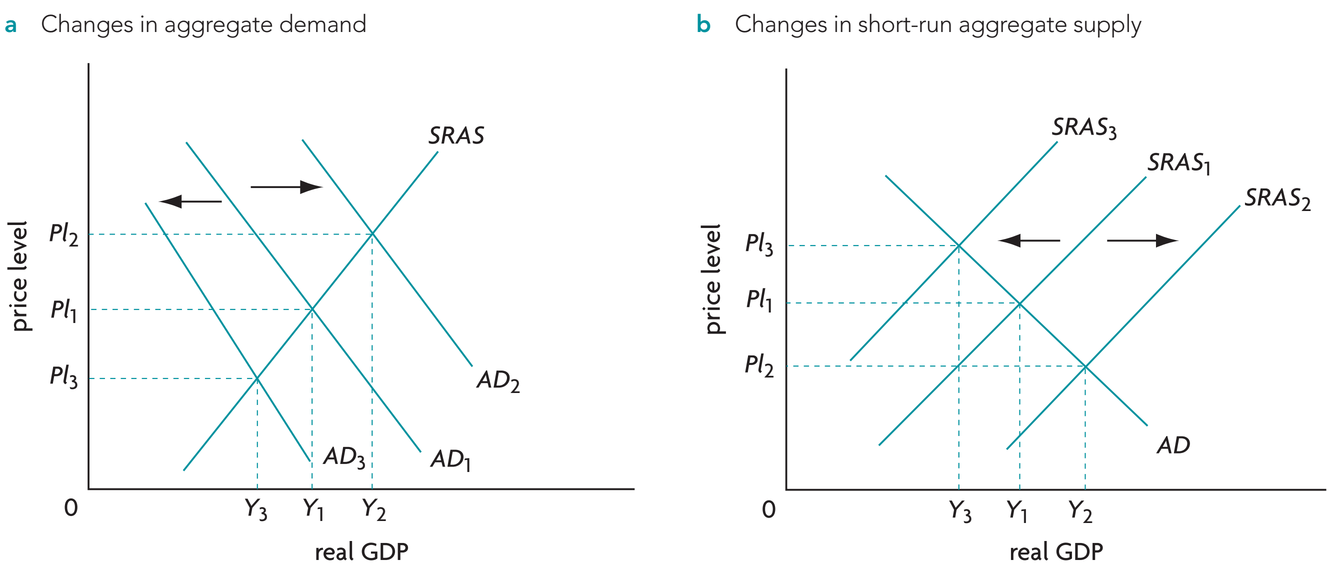


Figure 9.4: Impacts of changes in short-run macroeconomic equilibrium

9.3 Long-run aggregate supply and long-run equilibrium in the monetarist/new classical model

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the monetarist/new classical perspective on the long-run aggregate supply (*LRAS*) curve (AO2)
- explain that in the monetarist/new classical model macroeconomic equilibrium in the long run is determined at full employment (or potential) output (AO2)
- explain that when the economy is at long-run equilibrium (full employment equilibrium) unemployment is equal to the natural rate of unemployment (AO2)
- draw the *LRAS* curve and macroeconomic equilibrium in the long run (AO4)
- explain inflationary and deflationary (recessionary) gaps (AO2)
- explain how in the monetarist/new classical perspective the economy automatically adjusts to full employment output (AO2)

The monetarist/new classical model

The long-run aggregate supply curve and long-run equilibrium

This section examines the theoretical perspective of **monetarist/new classical** economists, which builds on the work of the classical economists of the 19th century. Both the monetarist/new classical and classical perspectives are based on the following key principles: the importance of the price mechanism in co-ordinating economic activities; the concept of competitive market equilibrium; and thinking about the economy as a harmonious system that automatically tends towards full employment. While economists generally accept these principles in the study of microeconomics, there is major disagreement over their relevance to the study of economics at the macro level. (See Chapter 1, Section 1.5 for a brief review.)

The monetarist/new classical approach to aggregate supply rests crucially on the distinction made earlier between the macroeconomic short run and long run. It examines what happens to aggregate supply when the economy moves into the long run, when all resource prices including wages change to match changes in the price level. The long-run supply relationship between the price level and aggregate output is referred to as **long-run aggregate supply (*LRAS*)**, shown graphically as the ***LRAS* curve**. The *LRAS* curve is vertical at potential GDP, also known as the full employment level of real GDP, Y_p, as shown in Figure 9.5. A vertical *LRAS* curve means that in the long run any change in *AD* results only in changes in the price level while the quantity of real GDP produced remains the same. The economy is in **long-run equilibrium** when the *AD* curve and the *SRAS* curve intersect at any point on the *LRAS* curve, seen in Figure 9.5.

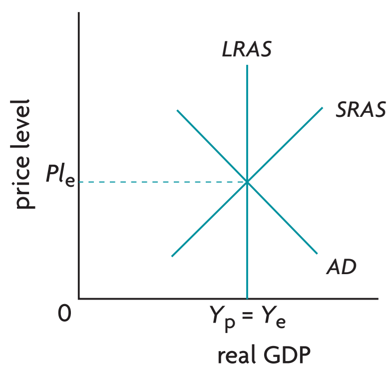


Figure 9.5: The *LRAS* curve and long-run equilibrium in the monetarist/new classical model

According to the monetarist/new classical perspective, the long-run aggregate supply (*LRAS*) curve is vertical at the full employment level of output, indicating that in the long run the economy produces potential GDP, which is independent of the price level. Long-run equilibrium occurs when the *SRAS* and *AD* curves intersect on the *LRAS* curve at the level of full employment or potential output.

Long-run equilibrium and the natural rate of unemployment

You may recall from Chapter 8 (Section 8.4), that when the economy produces at potential output, we say that the economy is experiencing 'full employment'. This is why the terms *full employment output* and *potential output* both refer to the output that is produced when the economy is at long-run equilibrium. Note that this is the level of output determined at the point where the *LRAS* curve is situated as you can see in Figure 9.5.

As we saw in Chapter 8, while we call this 'full employment output' the economy still has unemployed labour and other resources. The unemployment that exists when the economy is producing its full employment output is known as the natural rate of unemployment. This will become clearer to you in Chapter 10.

Why the *LRAS* curve is vertical

There is a very simple explanation for the vertical shape of the *LRAS* curve. Since wages (and other resource prices) are now *changing to match output price changes*, firms' costs of production remain constant even as the price level changes. Therefore, as the price level increases or decreases, *with constant real costs, firms' profits are also constant, and firms no longer have any incentive to increase or decrease their output levels*.

For example, say the price level increases. In the short run, with wages (and other input prices) constant, firms' profits increase, and firms therefore increase the quantity of output produced by moving upward along an upward-sloping *SRAS* curve. However, in the long run, wages (and other resource prices) also increase by the same amount. In effect, nothing has changed from the firms' point of view, and so they have no reason to increase the quantity of output they produce. Similarly, any price level decrease is fully matched by the same decrease in wages (and other resource prices), so that firms have no incentive to decrease the quantity of output produced.

Short-run equilibrium positions in relation to long-run equilibrium: deflationary (recessionary) gaps and inflationary gaps

As we have seen in Figure 9.5 when *AD* and *SRAS* intersect on the *LRAS* curve, there is long-run equilibrium. But what happens if *AD* and *SRAS* intersect at some other point that is not on the *LRAS* curve?

There are two such possibilities, shown in Figure 9.6(a) and (b). In all three diagrams Y_p represents potential output, or full employment output, which is given by the position of the *LRAS* curve on the horizontal axis. At Y_p, unemployment is equal to the natural rate of unemployment.

- **Figure 9.6(a): deflationary (recessionary) gap.** In part (a), equilibrium real GDP, Y_e, lies to the left of potential GDP, Y_p. When real GDP is less than potential GDP, the economy is experiencing a *deflationary gap* (also known as a *recessionary gap*), and unemployment is greater than the natural rate of unemployment. Why does this happen? The deflationary gap has been created because at the price level Pl_e, the amount of real GDP that the four components of aggregate demand want to buy is less than the economy's potential GDP. *There is not enough total demand in the economy* to make it worthwhile for firms to produce potential GDP. This also means that firms require less labour for their production; therefore, unemployment is greater than the natural rate of unemployment.
- **Figure 9.6(b): inflationary gap.** In part (b), equilibrium real GDP, Y_e, lies to the right of potential GDP, Y_p. When real GDP is larger than potential GDP, the economy is experiencing an *inflationary gap*, and unemployment is less than the natural rate of unemployment. An inflationary gap arises because with aggregate demand *AD*, the quantity of real GDP that the four components want to buy at the price level (Pl_e) is greater than the economy's potential output. *There is too much total demand in the economy*, and firms respond by producing a greater quantity of real GDP than potential GDP. To produce more output, firms' labour needs to increase, and unemployment falls to become less than the natural rate of unemployment.
- **Figure 9.6(c)**: **Full employment level of real GDP**, or **potential output**. Part (c) is the same as Figure 9.5, showing long-run equilibrium, where equilibrium real GDP is equal to full employment or potential GDP. When the economy is producing its potential GDP, unemployment is equal to the natural rate of unemployment and there is no deflationary or inflationary gap.

You can now see that the three states of the economy in Figure 9.6 correspond to the phases of the business cycle that we studied in Chapter 8: Y_e of Figure 9.6(a) corresponds to a point like e in Figure 8.4, where the economy is experiencing recession, unemployment is greater than the natural rate and actual GDP is less than potential GDP. Y_e of Figure 9.6(b) corresponds to a point like d in Figure 8.4, where unemployment is lower

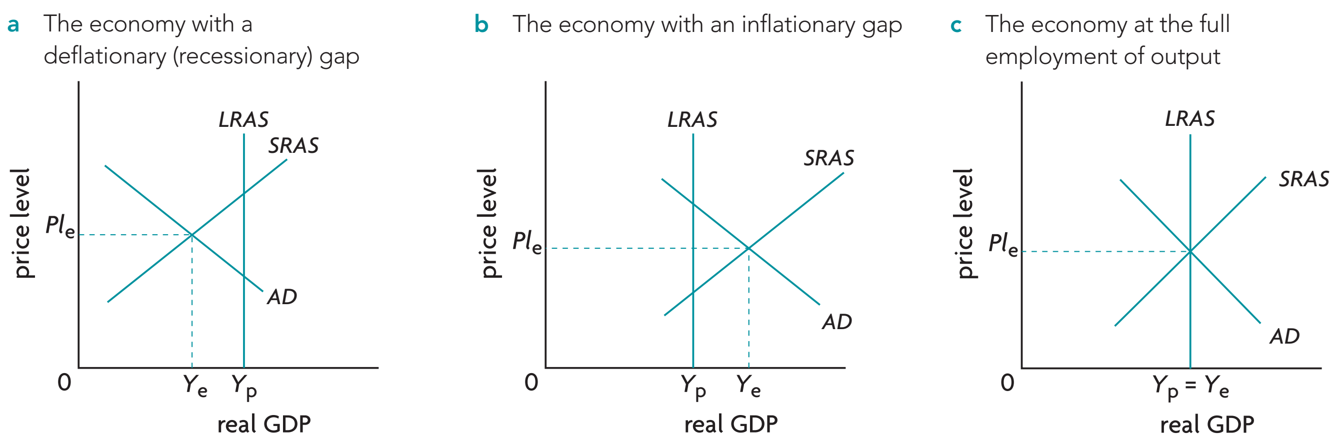


Figure 9.6: Deflationary (recessionary) and inflationary gaps in relation to potential output

than the natural rate and actual GDP is greater than potential GDP. Finally, Y_p of Figure 9.6(c) corresponds to points like a, b, and c in Figure 8.4, where the economy is producing actual GDP equal to potential GDP, with unemployment at its natural rate. Therefore, recessionary and inflationary gaps are two types of output gaps.

> Recessionary (deflationary) and inflationary gaps represent short-run equilibrium positions of the economy. A **deflationary (recessionary) gap** is a situation where real GDP is less than potential GDP (and unemployment is greater than the natural rate of unemployment) due to insufficient aggregate demand. An **inflationary gap** is a situation where real GDP is greater than potential GDP (and unemployment is smaller than the natural rate of unemployment) due to excess aggregate demand. When the economy is at its full employment equilibrium level of GDP, the *AD* curve intersects the *SRAS* curve at the level of potential GDP, and there is no deflationary or inflationary gap. This is the economy's **full employment level of output**, also known as **potential output**.

Shifts in *AD* or *SRAS* as possible causes of the business cycle

It is now a simple matter to consider the possible causes of the business cycle studied in Chapter 8. In Figure 9.7(a) and (b), the economy is initially at full employment equilibrium, producing potential output Y_p. In part (a), a fall in aggregate demand, shifting the *AD* curve leftward from AD_1 to AD_2 causes a recessionary gap. If the economy experiences an increase in aggregate demand, appearing as a rightward shift in the *AD* curve from AD_1 to AD_3, this causes an inflationary gap.

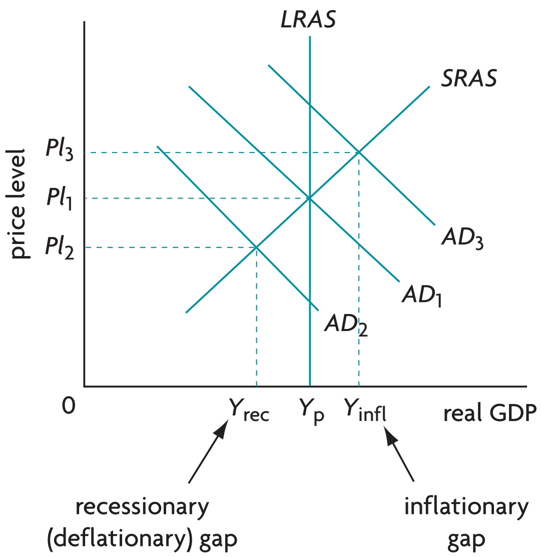
Shifts in the *SRAS* curve can also contribute to economic fluctuations.[4] In Figure 9.7(b), starting again from full employment equilibrium, a fall in *SRAS*, shifting $SRAS_1$ to $SRAS_2$, leads to an economic contraction, with real GDP falling to Y_2 and unemployment increasing. Note, however, that this contraction differs from the recessionary gap resulting from the fall in aggregate demand: the fall in aggregate supply leads to *an increase in the price level, along with a decrease in real GDP*. This special set of circumstances is especially undesirable for an economy, as it involves the appearance of two problems: recession (with unemployment) and a rising price level. This is known as *stagflation* (combining 'stagnation' with 'inflation'), a term coined in the 1970s. We will come back to this topic in Chapter 10.

An increase in *SRAS*, shifting $SRAS_1$ to $SRAS_3$ leads to an economic expansion as real GDP increases to Y_3 and unemployment falls. This expansion results in a falling price level, in contrast to the rising price level following an increase in aggregate demand.

[4] It may be noted that changes in aggregate supply can cause contractions and expansions; however, these are not called deflationary (recessionary) or inflationary gaps. The reason is that deflationary and inflationary gaps are defined in terms of the level of actual aggregate demand relative to the aggregate demand that is required to bring about full employment equilibrium. A deflationary gap is therefore caused by insufficient aggregate demand, and an inflationary gap by too much aggregate demand.

Most economists believe that *changes in aggregate demand are more frequent than changes in aggregate supply* as causes of the business cycle.

a Changes in aggregate demand



b Changes in short-run aggregate supply

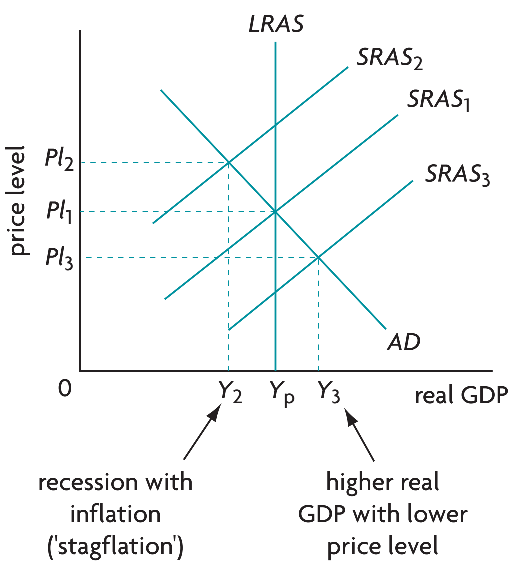


Figure 9.7: Possible causes of the business cycle

Automatic adjustment to full employment equilibrium at the level of potential GDP (or why inflationary and deflationary gaps cannot persist in the long run)

In our discussion above, we saw that inflationary and deflationary gaps are two possible short-run equilibrium positions of the economy where the equilibrium level of real GDP differs from potential GDP. If the *LRAS* curve is vertical at potential GDP, it follows that inflationary and deflationary gaps are only short-run phenomena that cannot persist in the long run. As soon as the economy moves into the long run, the gaps disappear, and the economy achieves full employment equilibrium.

To see how this occurs, consider Figure 9.8(a), where an economy is initially in long-run equilibrium at point a producing potential output, Y_p. A fall in aggregate demand from AD_1 to AD_2 causes the economy to move in the short run from point a to point b, where there arises a deflationary gap; at b, real GDP has fallen to Y_{def} and the price level has fallen from Pl_1 to Pl_2. However, the economy cannot remain there in the long run. In the long run, the fall in the price level is matched by a fall in wages (and falls in other resource prices), so the *SRAS* curve shifts to the right from $SRAS_1$ to $SRAS_2$ until the economy is back on the *LRAS* curve, at point c. *The assumption of wage and price flexibility in the long run has allowed the economy to automatically come back to its long-run equilibrium level of output*. The deflationary gap is eliminated, and the only thing that changes due to the fall in aggregate demand is the fall in the price level (from Pl_1 to Pl_3).[5]

In Figure 9.8(b) we see what happens if there is an increase in aggregate demand. Beginning from long-run equilibrium at point a, aggregate demand shifts from AD_1 to AD_2; in the short run the economy moves to point b, real GDP increases to Y_{infl} where there is an inflationary gap, and the price level increases from Pl_1 to Pl_2. However, the economy cannot remain at point b in the long run, because once wages (and other resource prices) increase to match the increase in the price level, *SRAS* shifts from $SRAS_1$ to $SRAS_2$, and the economy arrives at point c, which is once again on the *LRAS* curve. In the long run, the inflationary gap is eliminated and the only

[5] You may be wondering why wages will fall in the long run, thereby causing the shift in the *SRAS* curve that makes the economy move back to full employment equilibrium. The reason involves adjustments that take place in the labour market. As we know from our earlier discussion, if there is a recessionary gap, aggregate demand is weak and there is unemployment of labour that is greater than the natural rate of unemployment. This means that there is a surplus of labour in the labour market; in other words, the quantity of labour supplied is greater than the quantity of labour demanded. This creates pressures on wages to fall, so as to bring about a balance between the quantity of labour demanded by firms and the quantity supplied by workers. Therefore, wages fall in the long run, in order to eliminate the labour surplus, and when there is no longer any surplus labour, the economy reverts to long-run equilibrium through the shift in the *SRAS* curve.

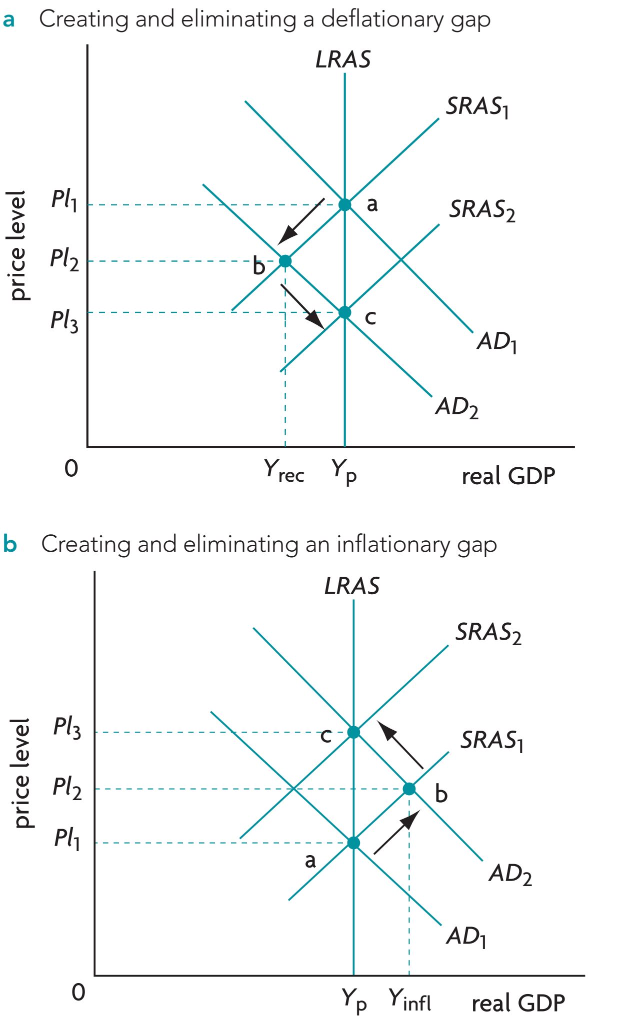


Figure 9.8: Automatic adjustment to long-run full employment equilibrium in the monetarist/new classical model

thing that changes after the increase in aggregate demand is the increase in the price level (to Pl_3).[6]

> In the monetarist/new classical perspective, recessionary (deflationary) and inflationary gaps are eliminated in the long run. This ensures that in the long run the *LRAS* curve is vertical at the level of potential GDP. The economy has a built-in tendency towards full employment equilibrium.

The move from point a to c in the long run in the case of a fall in *AD* that causes a deflationary gap (Figure 9.8(a)), and an increase in *AD* that causes an inflationary gap (Figure 9.8(b)), indicates the following important principle.

> In the monetarist/new classical perspective, changes in aggregate demand can have an influence on real GDP only in the short run; in the long run, the only impact of a change in aggregate demand is to change the price level, having no impact on real GDP, as this remains constant at the level of potential or full employment output (see also Figure 9.14(a)). Increases in aggregate demand in the long run are therefore inflationary (cause inflation).

TEST YOUR UNDERSTANDING 9.4

1. a Define the long run in macroeconomics and use a diagram to show the long-run aggregate supply (*LRAS*) curve.
   b Outline what the vertical shape of the *LRAS* curve tells us about the relationship between the price level and real GDP in the long run.
2. a Define and use a diagram to show long-run equilibrium in the *AD-AS* model (show the relationship between the *LRAS*, *SRAS* and *AD* curves).
   b Outline what the long-run equilibrium says about the level of unemployment.
3. Draw two diagrams, illustrating recessionary and inflationary gaps in relation to the *LRAS* curve.
4. Using the three possible states of the economy shown in Figure 9.6, explain the phases of the business cycle (refer to expansions, contractions, and potential output).

[6] When there is an inflationary gap, unemployment falls below the natural rate, and there is a shortage of labour in the labour market. Firms have a strong demand for labour (as well as other resources) and workers would like to negotiate higher wages because the price level has increased. In the long run, the wage is free to change in response to the forces of supply and demand, and moves upward to the point where quantity of labour demanded is brought into balance with quantity of labour supplied. When this occurs, the economy returns to long-run equilibrium through the shift in the *SRAS* curve.

CONTINUED

5 Using diagrams explain why inflationary or deflationary gaps (short-term fluctuations) cannot persist in the long run according to the monetarist/new classical perspective.

6 Assuming the economy begins from a position of full employment equilibrium, explain how each of the events listed in the question in Test your understanding 9.3 can contribute to short-term economic fluctuations.

9.4 Aggregate supply and equilibrium in the Keynesian model

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the Keynesian perspective of the aggregate supply (*AS*) curve (AO2)
- explain equilibrium in the Keynesian model (AO2)
- draw a diagram showing equilibrium in the Keynesian model (AO4)
- explain that in the Keynesian model deflationary/recessionary gaps may persist so that the equilibrium level of output may differ from the full employment level of output (AO2)

This section presents the theoretical model of Keynesian economists. Keynesian economists base their ideas on the work of John Maynard Keynes, one of the most famous economists of the 20th century, whose work in the first half of the century came to form the basis of modern macroeconomics. (See Chapter 1, Section 1.5 for an overview.) Keynes questioned the classical economists' view of the economic system as a harmonious system that automatically tends towards full employment, and showed that it is possible for economies to remain in a position of short-run equilibrium for long periods of time.

Getting stuck in the short run

Wage and price downward inflexibility

The *LRAS* curve in the monetarist/new classical model depends on the idea that all resource prices and product prices are fully flexible and respond to the forces of supply and demand. However, what if resource prices cannot fall, even over long periods of time? Keynesian economists argue that there is an asymmetry between wage changes in the upward and downward directions. Under conditions of an economic expansion and strong aggregate demand (rightward shifts in the *AD* curve causing an inflationary gap), with unemployment lower than the natural rate and a rising price level, wages quickly begin to move upward. Yet in a recessionary gap, where aggregate demand is weak and the economy is in recession with unemployment greater than the natural rate, wages do not fall easily, even over long periods of time, because of a variety of factors (such as labour contracts, minimum wage legislation; worker and union resistance to wage cuts; employer resistance due to morale).

Keynesian economists also argue that not only wages but also product prices do not fall easily, even if an economy is in a recessionary gap. In a recession, if wages will not go down, firms will avoid lowering their prices because that would reduce their profits. Furthermore, large oligopolistic firms may fear price wars; if one firm lowers its price, then others may lower theirs more aggressively in an effort to capture market shares, and then all the firms will be worse off. Such factors, it is argued, make prices unlikely to fall even in a recession.

The inability of the economy to move into the long run

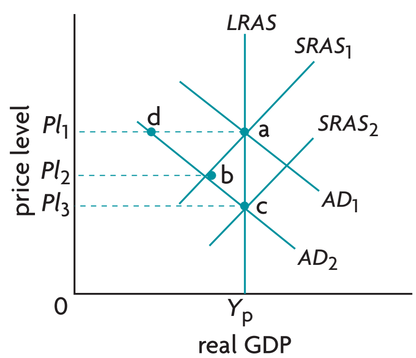
If wages and prices do not fall easily, this means the economy may get stuck in the short run. Consider Figure 9.9(a), which is similar to Figure 9.8(a). Beginning at point a where an economy is producing potential output Y_p, aggregate demand falls so the *AD* curve shifts from AD_1 to AD_2. The monetarist/new classical model predicts that the economy will move to point b in the short run, where there is a recessionary gap and the price level falls from Pl_1 to Pl_2; in the long run it will move to point c with Pl_3 and the economy is once again producing potential output Y_p.

However, if the price level cannot fall from Pl_1, the economy will move to point d on the new, lower, aggregate demand curve, AD_2. Even if the price level succeeds in falling to Pl_2, so the economy moves to point b, the economy may get stuck there if wages do not fall (remember that wages must fall for the *SRAS* curve to shift to $SRAS_2$ on the *LRAS* curve). It follows that if the price level cannot fall, or if wages cannot fall, the economy gets stuck in the short run, and is unable to move into the long run where it eliminates the recessionary gap.

This argument suggests that the *SRAS* curve has the shape shown in Figure 9.9(b). The horizontal part of the curve is based on the Keynesian idea that wages and prices do not move downward. Point d in Figure 9.9(a) corresponds roughly to point d in Figure 9.9(b). The economy is in a deflationary gap and may stay there indefinitely unless the government intervenes with specific policies.

In the Keynesian model, inflexible wages and prices in the downward direction mean that the economy cannot move into the long run when experiencing a deflationary gap. Inflexible wages and prices are shown graphically by a horizontal section of the Keynesian aggregate supply (*AS*) curve.

a The implications of downwardly inflexible wages and prices



b The Keynesian *AS* curve

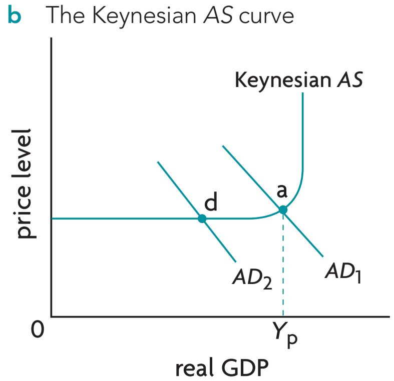


Figure 9.9: Keynesian analysis

Keynesians would not suggest that wages and prices can never fall. They would agree that if a recession or depression (which is a very severe recession) continues for a long enough time (perhaps years), wages and prices would eventually begin to fall. In the meantime a long-lasting recession would be very costly in terms of unemployment, low incomes and lost output. Therefore, it would be necessary for the government to intervene with active policies to help the economy come out of the recession.

The shape of the Keynesian aggregate supply curve

Figure 9.10 shows that the **Keynesian aggregate supply curve** has three sections. In section I, real GDP is low, and the price level remains constant as real GDP increases. In this range of real GDP, there is a lot of unemployment of resources and *spare capacity*. Spare capacity refers to the availability of resources including physical capital (machines, equipment, etc.) are labour that are not used. Firms can easily increase their output by employing the unemployed capital and other unemployed resources, without having to bid up wages and other resource prices. In section II, real GDP increases are accompanied by increases in the price level. The reason is that as output increases, so does employment of resources, and eventually bottlenecks in resource supplies begin to appear as there is no longer spare capacity in the economy. Wages and other resource prices begin to rise, which means that costs of production increase. The only way that firms will be induced to increase their output is if they can sell it at higher prices. Therefore, growing output leads to an increasing price level.

At output level Y_p, the economy has reached its full employment level of real GDP. This is also its potential output level, and unemployment has fallen to the point where it is now equal to the natural rate of

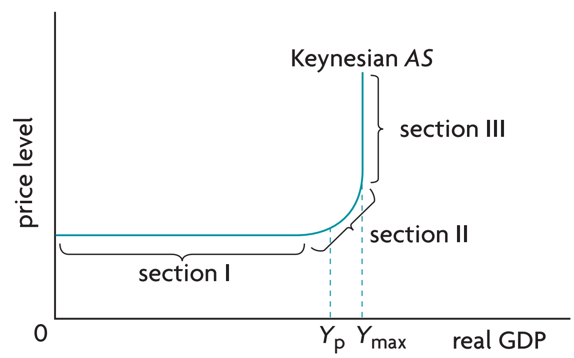


Figure 9.10: The Keynesian aggregate supply curve

unemployment. However, as we know, the natural rate of unemployment is not maximum employment, as unemployment can fall further, which is what happens when real GDP continues to increase beyond Y_p. Real GDP can continue to increase until it reaches section III.

In section III, the *AS* curve becomes vertical at Y_{max}, indicating that real GDP reaches a level beyond which it cannot increase anymore; at this point, the price level rises very rapidly. Real GDP can no longer increase because firms are using the maximum amount of labour and all other resources in the economy. Any efforts on the part of firms to increase their output only result in greater increases in the price level.

The three equilibrium states of the economy in the Keynesian model

Macroeconomic equilibrium in the Keynesian model is determined by the point where the *AD* curve intersects the Keynesian *AS* curve. This can occur at any level of real GDP. There are three equilibrium states of the economy, shown in Figure 9.11.

Figure 9.11(a) shows the *AD* curve intersecting the *AS* curve in its horizontal section, determining Y_e, which is less than Y_p (potential GDP), indicating a deflationary (recessionary) gap with unemployment greater than the natural rate. Aggregate demand is too weak to induce firms to produce at Y_p. In part (b), the economy is producing at Y_e, which is greater than Y_p, and is experiencing an inflationary gap. There is strong aggregate demand, unemployment has fallen below its natural rate, and as the economy approaches its maximum capacity, the price level has increased.

Part (c) shows the case where the economy has achieved full employment equilibrium, or potential output, at Y_p.

These three equilibrium states of the economy can be related to the business cycle (see Chapter 8): Y_e in Figure 9.11(a) corresponds to a point like e in Figure 8.4, where there is a deflationary gap; Y_e of Figure 9.11(b) corresponds to a point like d in Figure 8.4, where there is an inflationary gap; and Y_p of Figure 9.11(c) corresponds to points like a, b and c in Figure 8.4, where the economy's actual output is equal to its potential output.

It should be noted that 'potential output' and 'natural unemployment', which we have used to illustrate the three kinds of equilibrium, are actually *monetarist* concepts. On the other hand, inflationary and deflationary (recessionary) gaps are *Keynesian* concepts. As our analysis shows, the two models can usefully borrow concepts from each other in order to show how different real-world situations can be understood and interpreted differently depending on the theoretical approach used.

The Keynesian model arrives at some conclusions that differ significantly from the conclusions of the monetarist/new classical model. Very briefly, these are that:

- the economy in the Keynesian model can remain indefinitely stuck in a deflationary gap, unlike in the monetarist/new classical model where the economy automatically returns to full employment equilibrium
- increases in aggregate demand in the Keynesian model need not necessarily result in increases in the price level, unlike in the monetarist/new classical model where increases in aggregate demand always result in a higher price level.

As a result of these differences, the two models have very different policy recommendations about how to deal with some very important problems of the macroeconomy.

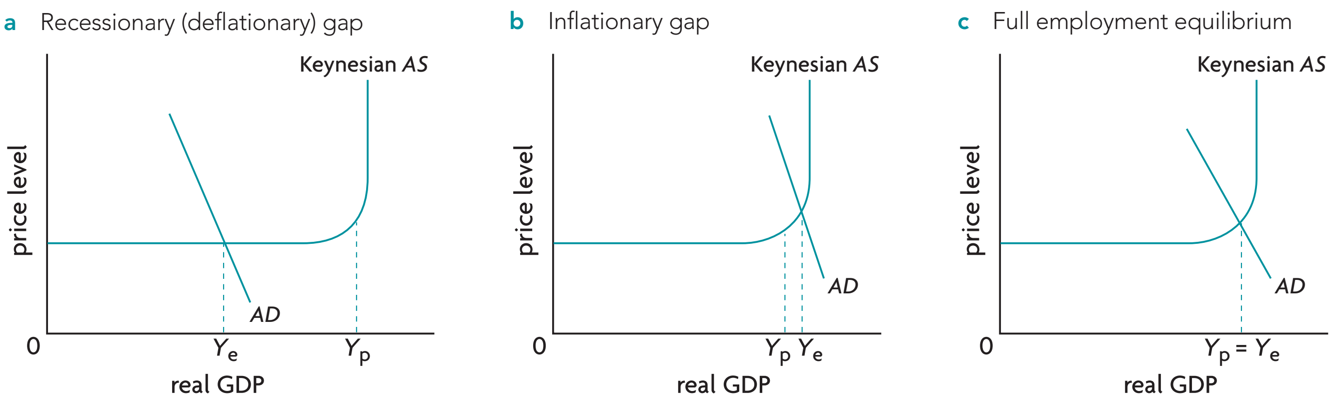


Figure 9.11: Three equilibrium states of the economy in the Keynesian model

TEST YOUR UNDERSTANDING 9.5

1 Define aggregate supply. Explain whether the meaning of this concept changes in the context of the *SRAS*, *LRAS* or Keynesian *AS* curves.

2 a Explain what it means for the shape of the aggregate supply curve if wages and prices are inflexible in the downward direction.

b Can the economy move into the long run?

c Outline what the horizontal section of the *AS* curve tells us about spare capacity in the economy.

3 a Use a diagram to show the Keynesian *AS* curve.

b Outline what the flat section of this curve indicates about the relationship between the price level and real GDP.

c Outline what the upward-sloping section indicates about this relationship.

d Outline what the vertical section indicates.

4 Using the Keynesian model and diagrams, show the three short-run equilibrium states of the economy, describing recessionary (deflationary) and inflationary gaps and their relationship to the full employment equilibrium position of the economy (potential output).

5 Using diagrams illustrating the Keynesian model, show and explain what happens to the equilibrium level of real GDP and the price level if aggregate demand shifts within

a the horizontal section of the Keynesian *AS* curve,

b the upward-sloping section of the Keynesian *AS* curve, and

c the vertical section of the Keynesian *AS* curve.

REAL WORLD FOCUS 9.1

The Italian economy contracts

In early 2019 business confidence in Italy fell to the lowest level in four years, suggesting that Italy's economic contraction which began the previous year may continue. Consumer confidence also fell. Falling real GDP for the last two quarters of the previous year meant that Italy was in recession. The economic downturn was the result of falling output in agriculture, forestry, fishing and industry. Net exports on the other hand increased, but not by enough to make up for the declines.

Source: *www.bloomberg.com; www.bbc.com*

Figure 9.12: Naples, Italy. Narrow street in the old part of the city

Applying your skills

1 Using *AD-AS* diagrams, explain the effect on Italy's real GDP of

a the drop in business and consumer confidence, and

b the increase in net exports.

2 Outline which components of aggregate demand were affected by the drop in business and consumer confidence.

3 Explain the likely position of the Italian economy in early 2019

a using a business cycle diagram,

CONTINUED

b using an *AD-AS* diagram with an *LRAS* curve, and

c using an *AD-AS* diagram with a Keynesian *AS* curve.

4 Outline which method of national income accounting allowed economists to determine

a the increase in Italy's net exports, and

b the decrease in output of the agriculture, forestry, fishing and industry sectors.

9.5 Shifting aggregate supply curves over the long term

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain that the *AS* curve shifts over the long run (monetarist/new classical *LRAS* model) or over the long term (Keynesian model) due to (AO2)
    - changes in quality and/or quality of factors of production
    - technological improvements
    - changes in efficiency
    - institutional changes
- draw diagrams showing shifts in the Keynesian *AS* and *LRAS* curves (AO4)

Changes in aggregate supply over the long term

Aggregate supply curve shifts in *AD-AS* models

So far, we have considered the *LRAS* and Keynesian *AS* curves in fixed, unchanging positions. Yet over time, these curves can shift. Each of these models shows a particular level of potential output, or the total quantity of goods and services produced by an economy when there is 'full employment' of its resources (meaning unemployment is equal to natural unemployment). Therefore, both curves shift to the right or to the left in response to factors that change potential output.

An increase in potential output signifies economic growth over the long term; a decrease signifies negative growth (or a fall in real output). Increases in potential output, shifting aggregate supply curves and long-term economic growth are illustrated in Figure 9.13. We will return to the topic of economic growth in Chapter 11.

Factors that change aggregate supply (shift *AS* curves) over the long term

The most important factors that cause increases in potential output and rightward shifts in the *LRAS* and Keynesian *AS* curves are the following:

- **Increases in quantities of the factors of production.** If the quantity of a factor of production increases, the *LRAS* curve and Keynesian *AS* curve shift to the right. For example, an increase in the quantity of physical capital, or the quantity of land (such as when there is a discovery of new oil reserves) means that the economy is capable of producing more real GDP. (If the quantity of factors of production decreases, the *LRAS* and *AS* curves shift to the left.)
- **Improvements in the quality of factors of production (resources).** Improvements in resource quality shift the *LRAS* and *AS* curves to the right. For example, greater levels of education, skills or health lead to an improvement in the quality of labour resources. More highly skilled and educated workers or healthier workers can produce more output than the same number of unskilled or less healthy workers.
- **Improvements in technology.** An improved technology of production means that the factors of production using it can produce more output, and the *AS* curves shift to the right. For example,

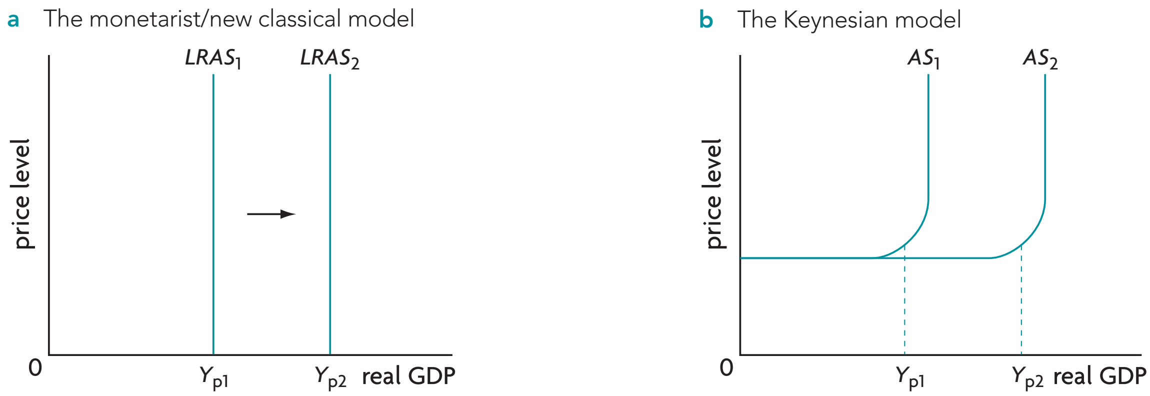


Figure 9.13: Increasing potential output, shifts in aggregate supply curves and long-term economic growth

workers who work with improved machines and equipment that have been produced as a result of technological innovations will be able to produce more output in the same amount of time.

- **Increases in efficiency.** When an economy increases its efficiency in production, it makes better use of its scarce resources, and can as a result produce a greater quantity of output. Therefore, potential output increases, and the *AS* curves shift to the right. (Decreases in efficiency would shift the *LRAS* and *AS* curves to the left.)
- **Institutional changes.** This point is related to efficiency in resource use because changes in institutions can sometimes have important effects on how efficiently scarce resources are used, and therefore on the quantity of output produced. For example, the degree of private ownership as opposed to public ownership of resources, the degree of competition in the economy, the degree and quality of government regulation of private sector activities, and the amount of bureaucracy can each affect the quantity of output produced (these points will be discussed in Chapter 13, under market-based supply-side policies).
- **Reductions in the natural rate of unemployment.** The natural rate of unemployment is the unemployment that is 'normal' or 'natural' for an economy when it is producing its 'full employment' level of output. It includes unemployed people who are in between jobs, who are retraining in order to become more employable, and others. The natural rate of unemployment differs from country to country and it can change over time. If it decreases, the economy is making better use of its resources, and can therefore produce a larger quantity of output. Therefore, potential output increases, and the *AS* curves shift to the right. (An increase in the natural rate of unemployment would result in a leftward shift in the *LRAS* and Keynesian *AS* curves.)

The relationship between the *SRAS* and *LRAS* curves in the monetarist/new classical model

If an economy is experiencing long-term economic growth, its *LRAS* curve will be shifting rightward, indicating increases in potential output. Yet over long periods of time, its *SRAS* curve will be shifting rightward as well. Any factor that shifts the *LRAS* curve must, over the long term, also shift the *SRAS* curve. This can be seen in Figure 11.3 (Chapter 11) in the discussion on economic growth.

Are there any factors that can shift the *SRAS* curve without shifting the *LRAS* curve? There are certain events with only a temporary effect on aggregate supply, and these can shift the *SRAS* curve for a short while, leaving the *LRAS* curve unchanged. Consider, for example, bad weather conditions that cause a drop in agricultural output. The *SRAS* curve shifts to the left for that season, but then moves back to the original position when the weather changes back to normal patterns; the *LRAS* curve remains unaffected. Changes in firms' costs of production, such as changes in wages, or changes in the prices of other key inputs (such as oil), may similarly affect only the *SRAS* curve. This applies to temporary changes that do not have a lasting impact on real GDP produced.

As an economy grows over time, it is likely that aggregate demand also increases. The reason is that many of the factors that cause the *LRAS* (and *SRAS*) curves to shift also cause the *AD* curve to shift. For example, increases in the quantity of physical capital, affecting *AS*, result from private and public investments, also shifting *AD*. We will return to this topic in Chapter 13.

TEST YOUR UNDERSTANDING 9.6

1 Illustrate diagrammatically the impacts on an economy's *LRAS* and the Keynesian *AS* curve of the following.

   a There is a widespread introduction of a new technology that increases labour productivity.

   b The government provides training programmes for workers to retrain and improve their skills.

   c A developing country receives large amounts of foreign aid, which allows it to purchase a large quantity of capital goods.

   d An extensive nationwide public health campaign undertaken by the government improves levels of health of the population.

   e The government introduces anti-monopoly legislation, reducing the market power of firms and increasing the economy's efficiency.

2 a Using diagrams and the concept of potential output, explain the relationship between long-term economic growth and the *LRAS* and Keynesian *AS* curves.

   b Identify some factors that can affect the *SRAS* curve but not the *LRAS* curve.

Illustrating the monetarist/new classical and Keynesian models

Figure 9.14 shows how the monetarist/new classical and Keynesian models relate to each other. Point a in both parts determines full employment equilibrium output, or potential GDP. Note that the *LRAS* curve in part (a) is not the same as the vertical section of the Keynesian *AS* curve, as the latter vertical section represents the maximum possible output that the economy can produce if it uses all its resources.

Point b in both parts represents a recessionary (deflationary) gap, which occurs due to low aggregate demand, given by AD_2 in parts (a) and (b). Point c in both parts represents an inflationary gap, which arises due to strong aggregate demand, given by AD_3.

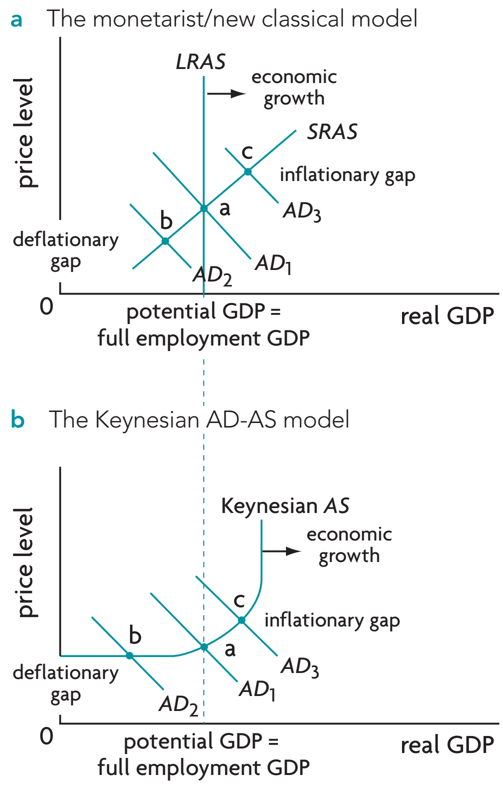


Figure 9.14: Comparing the monetarist/new classical and Keynesian models

Finally, economic growth is illustrated in both parts by the rightward pointing arrows. In part (a) it is represented by a rightward shift of the *LRAS* curve; in part (b) by a rightward shift of the Keynesian *AS* curve.

9.6 Implications of the Keynesian model and the monetarist/new classical model

LEARNING OBJECTIVES

After studying this section you will be able to:

- discuss the differing assumptions of the Keynesian and monetarist/new classical models and their implications for the economy and for policy (AO3)

Automatic self-correction versus persistence of deflationary gaps over long periods of time

Our study of Figure 9.8 showed that in the monetarist/new classical model inflationary and deflationary gaps are automatically corrected as long as resource prices, especially wages, are free to change as the price level changes. By contrast, one of the most important ideas arising from the Keynesian interpretation of the *AD–AS* model is that deflationary/recessionary gaps can persist over long periods of time. According to Keynes, this happens partly because of the inability of wages and prices to fall. In addition, the problem is caused by insufficient aggregate demand. Whenever aggregate demand intersects the horizontal section of the Keynesian *AS* curve, the economy is in a deflationary gap because aggregate demand is too low, and its four components are unable to buy enough output to make it worthwhile for firms to produce potential GDP. Therefore, equilibrium GDP is lower than potential GDP. In Figure 9.11(a), the equilibrium level of real GDP settles at Y_e, and can remain there indefinitely.

Keynesian analysis is therefore essentially a short-term analysis. This does not mean that Keynesian economists do not consider what happens over long periods of time; it means only that they do not accept the idea that the economy can move into what monetarist/new classical economists define as the long run (where there is full resource and product price flexibility).

> In contrast to the monetarist/new classical model, which automatically corrects deflationary/recessionary gaps by returning to full employment equilibrium, the Keynesian model shows that an economy can remain for long periods of time in an equilibrium where there is less than full employment (i.e. a deflationary/recessionary gap), caused by insufficient aggregate demand.

The two different models have important implications for economic policy. According to the monetarist/new classical model, governments should try to make markets work as freely as possible, so that wages and product prices can respond to the forces of demand and supply, without government interference in markets. By contrast, according to the Keynesian model, the government must intervene in the economy with specific measures to help it come out of the deflationary gap.

Increases in aggregate demand need not cause increases in the price level

Another important difference between the two models has to do with the effects of increases in aggregate demand on the price level. In the monetarist/new classical model, increases in aggregate demand always result in price-level increases. In the short run, as *AD* shifts to the right causing a movement along an upward-sloping *SRAS* curve, an increase in real GDP and an increase in the price level result, as can be seen in Figure 9.4(a). In the long run, increases in aggregate demand give rise only to increases in the price level, leaving real GDP unaffected, as in Figure 9.15(a). In the Keynesian model, when the economy is in the horizontal part of the *AS* curve, increases in aggregate demand lead to increases in real GDP without affecting the price level. This can be seen in Figure 9.15(b). It is only when the

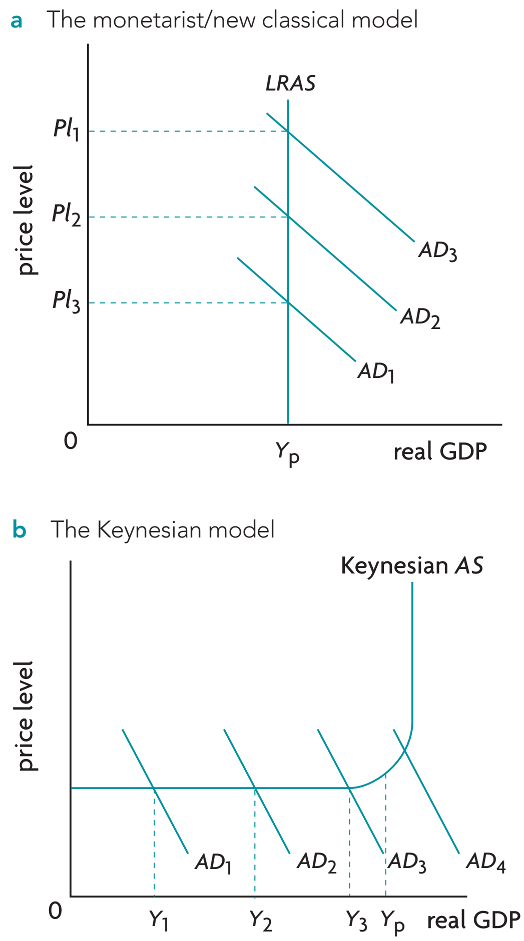


Figure 9.15: Effects of increases in aggregate demand on real GDP and the price level

Keynesian *AS* curve begins to slope upward, when it is close to the full employment level of output, that further increases in aggregate demand begin to result in changes in the price level as well. When the *AS* curve becomes vertical, increases in aggregate demand result in rapid price level increases while leaving real GDP unchanged.

In the Keynesian view increases in aggregate demand need not result in a higher price level. This is in contrast to the monetarist new/classical model where increases in aggregate demand always result in a higher price level.

Here too, there are differing implications for economic policy. As we will see in Chapter 10, rapid increases in the price level (or inflation) are undesirable. In the monetarist/new classical view, since increases in aggregate demand will always lead to increases in the price level, economic policy should focus on policies to achieve *long-run growth*, which are based on efforts to shift the *LRAS* curve to the right.

By contrast, in the Keynesian view, since increases in aggregate demand do not lead to price level increases when the economy is in a deflationary gap, policies focusing on increasing aggregate demand are not only harmless, but in fact are essential in order to both prevent and reduce the size of both deflationary and inflationary gaps.

TEST YOUR UNDERSTANDING 9.7

1 Compare and contrast the Keynesian and monetarist/new classical models regarding the ability of an economy to achieve full employment equilibrium on its own without any government intervention.

2 a Using the Keynesian model, explain when increases in aggregate demand can be expected to lead to increases in the price level (inflation) and when they are unlikely to do so.

  b Explain how the Keynesian model differs from the monetarist/new classical model in its prediction of rising price levels following an increase in aggregate demand.

3 Explain why use of the *LRAS* curve to account for economic growth leads to the policy implication that governments should focus on policies that try to influence the supply side of the economy.

4 Explain why use of the Keynesian three-section aggregate supply curve leads to the policy implication that governments should focus on the policies that try to influence the demand side of the economy.

THEORY OF KNOWLEDGE 9.1

Conflicting economic perspectives and the role of economists' political beliefs and ideology

The two perspectives we have studied in this chapter, the monetarist/new classical and Keynesian, are based on very different ways of viewing the economic world. The differences between the two are not just of theoretical interest; they have important implications for the real world, because each perspective provides very different policy recommendations to deal with macroeconomic problems.

The monetarist/new classical perspective

In the monetarist/new classical perspective, the economy is seen as a stable system that automatically tends towards long-run equilibrium where there is full employment at the natural rate of unemployment. This argument has important implications for the short-term fluctuations of the business cycle and long-term economic growth. Since short-term fluctuations (deflationary/recessionary and inflationary gaps) correct themselves automatically, there is no need for the government to do anything to correct them. Instead, the government must ensure that markets work as competitively as possible, so that all resource and product prices are able to rise or fall as required to allow the economy to settle at its point of long-run equilibrium, at the level of potential GDP.

CONTINUED

In fact, continues the argument, if governments do intervene with policies intended to correct short-term fluctuations, they may achieve the opposite of the intended results. Rather than reduce the size of fluctuations, they may make them bigger. Many monetarist/new classical economists believe that the departures of actual GDP from potential GDP that occur in real-world business cycles are as large as they are because of government intervention in the economy.

When it comes to promoting economic growth, aggregate demand cannot affect real GDP over the long run. If aggregate demand increases, it will only result in increasing price levels and inflation. Governments should therefore concentrate on policies that affect the supply side of the economy, which attempt to shift the *LRAS* curve to the right, with the objective of increasing real GDP without causing inflation.

The Keynesian perspective

In the Keynesian perspective, the economy is an unstable system because of repeating short-term fluctuations that cannot automatically correct themselves. Such fluctuations arise mainly due to changes in aggregate demand caused by spontaneous actions of firms and consumers. Keynes himself considered business cycle fluctuations to be caused mainly by changes in investment spending caused by changes in firms' expectations about the future. Optimism about the future increases investment spending, causing a rightward shift in the *AD* curve; pessimism decreases investment spending, leading to a leftward shift. Keynes referred to alternating waves of optimism and pessimism as 'animal spirits'.

In the Keynesian view, when there is a deflationary gap, there are many factors preventing the operation of market forces, and so wages and product prices do not fall easily even over long periods of time. This means the economy can remain in a less than full employment equilibrium (deflationary gap) for long periods. Therefore, there is an important role for government policy to play to restore full employment and raise real GDP to the level of potential GDP. Governments should focus on policies that increase aggregate demand when there is a deflationary gap, and decrease aggregate demand when there is an inflationary gap. Policies to influence aggregate demand are particularly important when aggregate demand is low.

Why does the debate persist?

Most economists today are unlikely to be purely 'monetarist'/'new classical' or purely 'Keynesian'. After decades of debate, many would argue that elements of both perspectives have some merit, and that policies attempting to influence both aggregate supply and aggregate demand are important in achieving the goals of reducing short-term fluctuations while promoting economic growth. Even so, most economists are still likely to side more with one perspective or the other. Why has the disagreement not been resolved after all these years? According to Mark Blaug, a prominent UK economist, there has been:

> *'... an unending series of efforts to produce a decisive empirical test of the Keynesian and monetarist view of the causes of economic fluctuations. A detached observer might be forgiven for thinking that this discussion has proved nothing but that empirical evidence is apparently incapable of making any economist change his mind ... But a closer look at the literature reveals ... a growing appreciation of the limitations of all the current statistical tests of the relative effectiveness of [government] policies... At the same time, it must be admitted that the persistence of this controversy, despite all the moves and countermoves in both camps, can only be explained in terms of certain deep-seated 'hard core' disagreements about the self-adjusting capacity of the private sector in mixed economies and, hence, the extent to which [government] policy is in fact stabilizing or destabilizing ... Once again, the debate between Keynesians and monetarists shows that economists (like all other scientists) will characteristically defend the core of their beliefs from the threat of observed anomalies...'*[7]

[7] Mark Blaug (1980) *The Methodology of Economics*, Cambridge University Press, pp. 217 and 221.

CONTINUED

Blaug is suggesting that the controversy persists because economists have different beliefs. What kind of beliefs could these be? On a general level, they must be beliefs about the superiority of one perspective over the other. However, this begs the question, where did these beliefs come from, and how can they be justified? Certainly not by the scientific method, based on empirical testing, since as Blaug clearly tells us, it has not been possible for an empirical test to falsify one or the other perspective based on the effectiveness of their policy recommendations. Therefore, very likely, these are beliefs that come from outside the realm of social scientific thinking, which may be political and ideological beliefs stemming from personal values.

Most economists do not deny the role of values and ideology in economics. Nobel Prize-winning economist, Robert Solow, writes the following:

> *'Social scientists, like everyone else, have class interests, ideological commitments, and values of all kinds. But all social science research, unlike research on the strength of materials or the structure of the haemoglobin molecule, lies very close to the content of those ideologies, interests, and values. Whether the social scientist wills it or knows it, perhaps even if he fights it, his choice of research problem, the questions he asks, the questions he doesn't ask, his analytical framework, the very words he uses, are all likely to be, in some measure, a reflection of his interests, ideologies and values.'*[8]

Thinking points

- Do you agree with Solow that it is very likely that personal value judgements influence economists' choices between alternative theories (the choice of 'analytical framework') and more generally their work as social scientists?
- Is the effective use of the scientific method influenced by economists' personal beliefs and ideologies?
- Do the social sciences, and economics in particular, differ from the natural sciences by having political beliefs and ideologies influence thinking?
- What kind of political beliefs and ideologies do you think are likely to be linked with (a) the monetarist/new classical perspective, and (b) the Keynesian perspective?

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Research an economy that has recently been (or is currently) in a recession.

2 Identify the causes of the recession and try to link them with what you have learned about the factors that cause aggregate demand and/or short-run aggregate supply to shift.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

[8] Robert M. Solow (1996) 'Science and ideology in economics' in D. M. Hausman, *The Philosophy of Economics*, Cambridge University Press.

Chapter 10

Macroeconomic objectives I:

Low unemployment, low and stable rate of inflation

BEFORE YOU START

- Unemployment occurs when people who are looking for a job cannot find one. Why do you think people become unemployed?
- What problems are associated with unemployment on a personal and social level? Can you think of some actions governments can take to help unemployed people become employed?
- Prices of goods and services tend to rise over time. This is known as 'inflation'. Why do you think inflation occurs over time?

This chapter is concerned with two important macroeconomic objectives: low unemployment and low and stable rate of inflation. Both of these are closely related to achieving potential output, shown in the business cycle diagram in Chapter 8; see Figure 8.4.

Photo: Bangkok, Thailand. Beggar sitting outdoors

10.1 Low unemployment

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain how unemployment and the unemployment rate are calculated (AO2)
- calculate the unemployment rate (AO4)
- explain the difficulties of measuring unemployment (AO2)
- explain and discuss the economic, personal and social costs of unemployment (AO3)[1]
- explain the causes of unemployment: cyclical, structural, seasonal and frictional (AO2)
- draw diagrams showing (AO4)
    - a fall in labour demand for a particular market or geographical area
    - unemployment resulting from a minimum wage
    - deflationary gap showing cyclical unemployment
- explain that the natural rate of unemployment is the sum of structural, seasonal and frictional unemployment (AO2)

Unemployment and its measurement

Unemployment, in a general sense, refers to idle, or not fully used resources. When economists use the terms 'unemployment' (or 'employment') on their own, they usually refer to unemployment or employment of labour. If they want to refer to another factor of production, they refer to it explicitly, such as 'unemployment of capital resources', or more generally, 'unemployment of resources'.

Our discussion in this chapter will focus on the economy's labour resources. Unemployment is defined as follows:

> **Unemployment** refers to people of working age who are actively looking for a job but who are not employed.

A closely related term is *underemployment*, referring to people of working age with part-time jobs when they would rather work full time, or with jobs that do not make full use of their skills and education. Examples include people who work fewer hours per week than they would like, or trained individuals, such as engineers, economists, or computer analysts, who work as taxi drivers, or waiters or waitresses, or anything else unrelated to their profession, when they would rather have a job in their profession.

Both unemployment and underemployment mean that an economy is wasting scarce resources by not using them fully. In the case of unemployment this is obvious. With underemployment, working at a job other than in one's profession also involves resource 'waste', because some resources that were used for training and education are wasted when people are forced to work at a job that does not make use of their skills.

Calculating unemployment: the unemployment rate

The *labour force* is defined as the number of people who are employed (working) plus the number of people of working age who are unemployed (not working but seeking work). The labour force is actually a fraction of the total population of a country, because it excludes children, retired persons, adult students, all people who cannot work because of illness or disability, as well as all people who do not want to work.

Unemployment can be measured as a number or percentage:

- As a number, unemployment is the total number of unemployed persons in the economy, i.e. all persons of working age who are actively seeking work but are not employed.
- As a percentage, unemployment is called the *unemployment rate*, defined as

[1] This point appears as AO2 in the syllabus, however it becomes AO3 in the point entitled 'Relative costs of unemployment versus inflation' considered later in this chapter.

$$\textbf{unemployment rate} = \frac{\text{number of unemployed}}{\text{labour force}} \times 100$$

For example, if the unemployment rate in an economy is 6%, this means that six out of every 100 people in the labour force are unemployed.

Suppose there is a population of 35.5 million people, of whom 17.3 million are in the labour force, 1.5 million work part time though they would rather work full time, and 1.4 million are looking for work but cannot find any. What is the unemployment rate?

The unemployment rate is $\frac{1.4}{17.3} \times 100 = 8.1\%$, which is the number looking for work unable to find any divided by the size of the labour force, times 100. Note that we ignore the size of the total population (it is irrelevant), as well as the number of people who are working part-time (they are considered to be employed).

Underemployment can similarly be measured as a number or as a percentage. If the underemployment rate is 15%, this means that 15 out of every 100 people in the labour force are underemployed.

Difficulties in measuring unemployment

The unemployment rate is one of the most widely reported measures of economic activity, used extensively as an indicator of economic performance. Yet it is actually difficult to obtain an accurate measurement of unemployment.

Official statistics often underestimate true unemployment because of *hidden unemployment*, arising from the following:

- Unemployment figures include unemployed persons who are actively looking for work. This excludes 'discouraged workers', who are unemployed workers who gave up looking for a job because, after trying unsuccessfully to find work for some time, they became discouraged and stopped searching. These people in effect drop out of the labour force.
- Unemployment figures do not make a distinction between full-time and part-time employment, and count people with part-time jobs as having full-time jobs though in fact they are underemployed.
- Unemployment figures make no distinction on the type of work done. If a highly trained person works as a waiter, this counts as fully employed.
- Unemployment figures do not include people on retraining programmes who previously lost their jobs, as well as people who retire early although they would rather be working.

In addition, official statistics may overestimate true unemployment, because:

- unemployment figures do not include people working in the underground economy (or informal economy). This is the portion of the economy that is unregistered, legally unregulated and not reported to tax authorities. Some people may be officially registered as unemployed, yet they may be working in an unreported (underground) activity.

A further disadvantage of the national unemployment rate (calculated for an entire nation) is that it is an average over the entire population, and therefore does not account for differences in unemployment that often arise among different *population groups* in a society. Within a national population, unemployment may differ by:

- region – regions with declining industries may have higher unemployment rates than other regions
- gender – women sometimes face higher unemployment rates than men
- ethnic groups – some ethnic groups may be disadvantaged due to discrimination, or due to lower levels of education and training
- age – youth unemployment (usually referring to persons under the age of 25) often face higher unemployment rates than older population groups, often due to lower skill levels; people who are ageing also sometimes face higher unemployment rates as employers may be less willing to employ them
- occupation and educational attainment – people who are relatively less skilled may have higher unemployment rates than more skilled workers (though in some countries higher unemployment rates may be found among highly educated groups).

Costs of unemployment

Unemployment of labour is one of the most important economic concerns to countries around the world. Reduction of unemployment is a key objective of governments everywhere, as its presence has major economic and social consequences.

Economic costs

Unemployment has the following economic consequences:

- **A loss of real output (real GDP).** Since fewer people work than are available to work, the amount of output produced is less than the level the economy is capable of producing. This is why unemployment means that an economy finds itself somewhere inside its production possibility curve (*PPC;* see Chapter 1, Section 1.3), producing a lower level of output than it is capable of producing.

- **A loss of income for unemployed workers.** People who are unemployed do not have an income from work. Even if they receive unemployment benefits, they are likely to be worse off financially than if they had been working.
- **A loss of tax revenue for the government.** Since unemployed people do not have income from work, they do not pay income taxes; this results in less tax revenue for the government.
- **Costs to the government of unemployment benefits.** If the government pays unemployment benefits to unemployed workers, the greater the unemployment, the larger the unemployment benefits that must be paid, and the less tax revenue left over to pay for important government-provided goods and services such as public goods and merit goods.
- **Costs to the government of dealing with social problems resulting from unemployment.** The social problems that arise from unemployment (noted below) often require government funds to be appropriately dealt with.
- **Larger budget deficit or smaller budget surplus.** A government budget deficit occurs when tax revenues are less than government expenditures, while a budget surplus is the opposite, involving greater tax revenues than expenditures. Unemployment leads to a loss of tax revenue for the government as we have seen, but at the same time greater expenditures for unemployment benefits as well as social problems due to unemployment. As expenditures rise while tax revenues fall, a budget surplus will become smaller while a budget deficit will become larger, in turn leading to more government debt.
- **More unequal distribution of income.** Some people (the unemployed) become poorer while others (the employed) are able to maintain their income levels. Since certain population groups (ethnic groups, regional groups, etc., discussed earlier) may be affected more by unemployment than others, the effects of increasing income inequalities and resulting poverty are often concentrated among population groups who are more disadvantaged to begin with. If unemployment is high or tends to persist over long periods of time, this may lead to increased social tensions and social unrest.
- **Unemployed people may have difficulties finding work in the future.** When people remain out of work for long periods, they may not find work easily at a later time in the future. This can happen because the unemployed workers may partly lose their skills due to not working for a long time, or because in the meantime new skills may be required that workers have not been able to keep up with, or because firms have found ways to manage with fewer workers. This process is known as *hysteresis* (from the Greek word *υστέρηση* meaning 'delay' or 'lagging behind something', in this case the lagging behind of employment). Hysteresis suggests that high unemployment rates in the present may mean continued high unemployment rates in the future, even when economic conditions become more favourable.

Personal and social costs

Unemployment has the following personal and social consequences:

- **Personal problems.** Being unemployed and unable to secure a job involves a loss of income, increased indebtedness as people must borrow to survive, as well as loss of self-esteem. All these factors cause great psychological stress, sometimes resulting in lower levels of health, family tensions, family breakdown and even suicide.
- **Greater social problems.** High rates of unemployment, particularly when they are unequally distributed for the reasons noted earlier, can lead to serious social problems, including increased crime and violence, drug use and homelessness, arising from growing poverty.

TEST YOUR UNDERSTANDING 10.1

1. Define unemployment and explain how it differs from underemployment.
2. Outline how we measure the unemployment rate.
3. Explain why unemployment figures are not usually accurate.
4. Identify some of the economic and social consequences of unemployment.
5. Calculate the unemployment rate in an economy with a population of 57.7 million people, of which the labour force is 62%, and the number of employed are 32.9 million.

Types and causes of unemployment

We will examine four types of unemployment: structural, frictional, seasonal and cyclical.

Structural unemployment

Structural unemployment occurs as a result of changes in demand for particular types of labour skills, changes in the geographical location of industries and therefore jobs, and labour market rigidities.

Changes in demand for particular labour skills

The demand for particular types of labour skills changes over time. This may be the result of technological change, which often leads to a need for new types of skills, while the demand for other skills falls. For example, computer technology, the introduction of automated teller machines (ATMs), and electrical relays and digital switching technology greatly reduced the need for typists, bank tellers and telephone operators, while increasing the need for workers with computer literacy and computer programming and other skills. There are growing concerns that automation (the introduction of automatic machines in manufacturing) may increasingly lead to job losses and therefore to more of this type of unemployment. According to a study by Oxford Economics, the introduction of robots around the world may replace 20 million manufacturing jobs by 2030. Those who lose their jobs will likely find jobs in other sectors including transport, construction, maintenance and office work. The study notes that areas where workers tend to have lower skills will be strongly affected.[2]

In addition, changes in demand for labour skills may occur because of changes in the structure of the economy, leading to some growing industries and some declining industries. Workers who lose their jobs in declining industries may not have the necessary skills to work in growing industries, and become structurally unemployed. For example, as the agricultural sector declines in relative importance and the manufacturing and services sectors grow, agricultural workers may lose their jobs. Workers lacking the necessary skills to work in industry or services may become structurally unemployed. (This type of structural change was explained in terms of income elasticity of demand (*YED*); see Chapter 3.)

These kinds of changes lead to *mismatches* between labour skills demanded by employers and labour skills supplied by workers. Such mismatches cause structural unemployment.

Changes in the geographical location of jobs

When a large firm or even an industry moves its physical location from one region to another, there is a fall in demand for labour in one region and an increase in the region where it relocates. (The same problem could arise if a large firm or industry closes down.) If people cannot move to economically expanding regions, they may become structurally unemployed. Sometimes firms relocate to foreign countries, increasing the overall structural unemployment within a country. Once again, the result will be a *mismatch* between labour demanded and labour supplied within a geographical region (or country).

Using a diagram to show structural unemployment arising from mismatches between labour demand and labour supply

Structural unemployment arising from mismatches between labour demand and supply can be shown in Figure 10.1(a), showing the labour market for a particular industry (or market) or for a particular geographical area. The vertical axis measures the wage (the price of labour) while the horizontal axis measures the quantity of labour. The demand curve shows the quantity of labour firms are willing and able to hire at each wage, and the supply curve shows the quantity of labour workers supply at each wage. The initial equilibrium is determined by S and D_1 at W_1 and Q_1.

Suppose now that due to technological change that reduces demand for some labour skills, or due to a change in the structure of the economy, or due to the move of an industry to another geographical area, there is a fall in the demand for labour, shown by the shift from D_1 to D_2. If market forces worked perfectly, the problem of structural unemployment would be solved. At the lower wage W_2, only Q_2 workers would want to work, so at that lower wage there would not be any excess supply of labour. However in practice, wages do not fall easily over short periods of time, and if they were to fall they would need a long time to do so. This means that if the wage remains at W_1, at least for the foreseeable future, this gives rise to an excess supply of labour that corresponds to structural unemployment created by the fall in the demand for

[2] **Source:** https://www.bbc.com/news/business-48760799

a Mismatches between labour demand and labour supply: falling demand for labour

b Minimum wage legislation and labour union activities lead to higher than equilibrium wages and lower quantity of labour demanded

c Labour market rigidities lead to an increase in costs of production (supply shifts to the left), causing a fall in Q produced; employers hire fewer workers

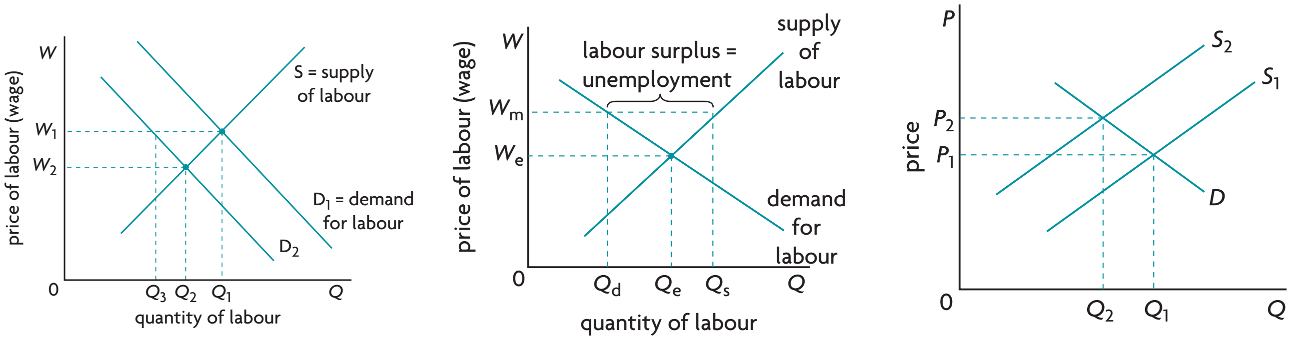


Figure 10.1: Structural unemployment

labour. Even if the wage falls a little, there would still be an excess supply of labour as long as it does not drop to the level of W_2.[3]

Labour market rigidities

Labour market rigidities are factors preventing the forces of supply and demand from operating in the labour market. They include:

- **minimum wage legislation**, which leads to higher than equilibrium wages
- **labour union activities and wage bargaining with employers**, resulting in higher than equilibrium wages
- **employment protection laws**, which make it costly for firms to fire workers (because they must pay compensation), thus making firms more cautious about hiring
- **generous unemployment benefits**, which increase the attractiveness of remaining unemployed and reduce the incentives to work.

Although economists do not always agree on the effects of these factors on unemployment, many argue that they are responsible for higher unemployment rates in countries with strong labour protection systems (such as in Europe) compared to countries with weaker labour protection systems (such as the United States).

Using diagrams to show structural unemployment arising from labour market rigidities

Unemployment arising from minimum wages and labour union activities leading to higher than equilibrium wages, can be shown in a labour market diagram, as in Figure 10.1(b). The higher than equilibrium wage, W_m, results in unemployment of labour equal to $Q_s - Q_d$. You may note that this diagram is the same as Figure 4.9 (Chapter 4) showing a price floor applied in the labour market.

Unemployment arising from minimum wage legislation, labour union activities and employment protection laws can also be shown indirectly, through a product supply and demand diagram, as in Figure 10.1(c). Higher than equilibrium wages and employment protection lead to higher costs of production for firms, causing the firm supply curve to shift to the left from S_1 to S_2, leading to a smaller quantity of output produced, Q_2 instead of Q_1. Firms therefore hire a smaller quantity of labour, and this contributes to structural unemployment of these types.

Structural unemployment (of all types) is a serious type of unemployment because it tends to be long term. A certain amount of structural unemployment is unavoidable in any dynamic, growing economy, and is therefore considered to be part of 'natural unemployment'. However, this does not mean it cannot be lowered. There are many

[3] Some economists make use of a different type of labour market diagram, which shows two labour supply curves: one for the labour force, and another representing the willingness of workers to take on jobs at different wage levels; the horizontal difference between the two supply curves at the market clearing wage represents natural unemployment (of which structural is the most important). This diagram presents natural (hence structural) unemployment as being wholly voluntary; jobs exist, but workers choose not to take them because they do not want them. This is a highly inaccurate representation of structural unemployment, which is that people want to work but cannot find jobs.

policies governments can pursue to reduce it, including measures encouraging workers to retrain and obtain new skills, and to relocate (move) to areas with greater employment opportunities; providing incentives to firms to hire structurally unemployed workers; and measures to reduce labour market rigidities. These policies will be discussed in Chapter 13.

Frictional unemployment

Frictional unemployment occurs when workers are between jobs. Workers may leave their job because they have been fired, or because they are in search of a better job, or they may be waiting to start a new job. Frictional unemployment tends to be short term, and does not involve a lack of skills that are in demand. It is therefore less serious than structural unemployment.

A certain amount of frictional unemployment is inevitable in any growing, changing economy, where some industries expand while others contract, some firms grow faster than others, and workers seek to advance their income and professional positions. An important cause of frictional unemployment is incomplete information between employers and workers regarding job vacancies and required qualifications. Imagine 100 job vacancies and 100 job applicants who have exactly the right job qualifications. Because of incomplete

REAL WORLD FOCUS 10.1

The textile industry in Naoussa

Naoussa is a city in northern Greece with a centuries-old textiles industry, based on highly labour-intensive production methods. When the markets of neighbouring transition economies opened up in the 1990s, Greek firms found it profitable to relocate to countries including Albania, Bulgaria, North Macedonia and Romania, on account of their far lower labour costs. The Greek government introduced legislation intended to lower labour costs (easier firing rules, extension of over-time work), yet Greek firms used some of these provisions to reduce their local workforce and move abroad.

In 2005, the removal of trade barriers on imports of Chinese textiles (according to WTO rules; see Chapter 15) led to a huge increase in Chinese textiles in the Greek market, forcing many Greek textile firms to close down as they were unable to compete with the lower-cost Chinese imports.

Figure 10.2: Greek textile design with olive branches

Naoussa was one of the areas most strongly affected. The combination of firm relocations and firm closures led to the loss of tens of thousands of jobs. In 2005, unemployment in Naoussa was estimated to have reached a record 35–40%.

Source: *Adapted from Mary Lembessi, 'Clothing exports dealt a blow' in Kathimerini, 23 September 2005; European industrial relations observatory on-line', Measure adopted in support of redundant textiles workers', 2 October 2006.*

Applying your skills

1 Identify the kind of unemployment was Naoussa experiencing by 2005.

2 Use a demand and supply diagram to explain how this type of unemployment came about.

3 **a** Use a demand and supply diagram to show the effects of legislation intending to lower labour costs in the textile industry.

b Why do you think this legislation was ineffective in keeping Greek firms from relocating?

information, it takes time for the right applicants to get matched up with the right jobs. Therefore, frictional unemployment is also part of natural unemployment.

Measures to deal with frictional unemployment aim at reducing the time that a worker spends in between jobs and improving information flows between workers and employers (see Chapter 13).

Seasonal unemployment

Seasonal unemployment occurs when the demand for labour in certain industries changes on a seasonal basis because of variations in needs. Farm workers experience seasonal unemployment because they are hired during peak harvesting seasons and laid off for the rest of the year. The same applies to lifeguards and gardeners, who are mostly in demand during summer months, people working in the tourist industry, which varies from season to season, shop assistants, who are in greater demand during peak selling months, and many others.

Some seasonal unemployment is unavoidable in any economy, as there will always be some industries with seasonal variations in labour demand. Therefore, seasonal unemployment is also part of natural unemployment.

Structural, frictional and seasonal unemployment: the natural rate of unemployment

As you know from Chapters 8 and 9, when the economy produces at full employment output, or potential output, it has unemployment equal to the natural rate of unemployment. We are now in a position to define this more accurately. The natural rate of unemployment is equal to the sum of structural, frictional plus seasonal unemployment. Another way of saying this is that when an economy has 'full employment', it actually has unemployment equal to the sum of structural, frictional and seasonal unemployment. Theory of knowledge 10.1 below examines this peculiar terminology.

Cyclical (demand-deficient) unemployment

We have seen what types of unemployment exist when an economy is producing at its potential or full employment level of output. What about unemployment arising when the economy produces less than its potential output? Unemployment now consists of additional unemployment, over and above the natural rate, which is known as *cyclical unemployment*.

Cyclical unemployment, as the term suggests, occurs during the downturns of the business cycle, when the economy is in a deflationary/recessionary gap. The downturn is seen as arising from declining or low aggregate demand (AD), and so is also known as **demand-deficient unemployment**. As real GDP falls due to a fall in AD, unemployment increases because firms lay off workers. In the upturn of the business cycle, as real GDP increases, the deflationary/recessionary gap becomes smaller and cyclical unemployment falls. When the economy produces real GDP at the level of potential output, there is no longer any cyclical unemployment, it is equal to zero.

Although cyclical unemployment is a Keynesian concept, it can be illustrated by use of both the monetarist/new classical and Keynesian versions of the AD-AS model, shown in Figure 10.3. In both parts, the economy is initially producing potential output Y_p, with zero cyclical unemployment. A fall in aggregate demand, causing AD_1 to shift to AD_2, creates a deflationary/recessionary gap as real output falls to Y_{rec}. At Y_{rec}, the new unemployment created is cyclical unemployment.

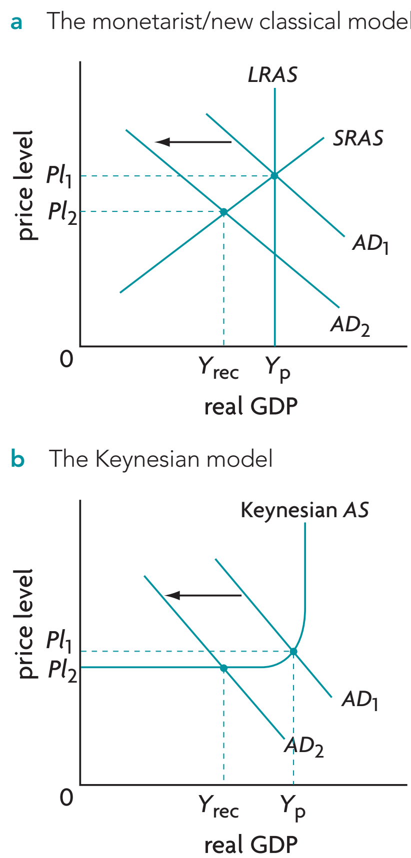


Figure 10.3: Cyclical unemployment

Since cyclical unemployment arises from a deficiency of aggregate demand, measures to reduce this unemployment involve the use of government policies to increase aggregate demand, and eliminate the recessionary gap (see Chapter 13).

The four types of unemployment in relation to the *AD-AS* model

The four types of unemployment are shown in relation to the *AD-AS* model in Figure 10.4. (For simplicity, the monetarist/new classical model is shown.) At output Y_p, real GDP is equal to potential or full employment GDP, where there is unemployment equal to the natural rate, or the sum of structural, frictional and seasonal unemployment, and cyclical unemployment is equal to zero.

If GDP falls to any level less than Y_p, there is a deflationary/recessionary gap, and unemployment increases so that in addition to structural plus frictional plus seasonal unemployment there is also cyclical (demand-deficient) unemployment. If GDP increases to any level greater than Y_p, there is an inflationary gap, and unemployment falls below the natural rate of unemployment. This means that some workers who were structurally, frictionally or seasonally unemployed now find jobs. However, these jobs tend to be of a short duration, because the economy does not usually remain in an inflationary gap indefinitely. The government is likely to step in with policies (that we will study in Chapter 13) to bring the economy back to output level Y_p, where unemployment will once again reach the natural rate.

Whereas it is a simple matter to distinguish between the four types of unemployment on a theoretical level, in the real world it can be very difficult to identify and distinguish between the different types of unemployment. The labour market is in a continuous state of change, with some workers quitting their jobs, others being fired, with some unemployed workers waiting for an appropriate job and others retraining for a new job, with some firms expanding, others contracting, and with some people newly entering the labour force and others leaving. The uncertainties surrounding the causes of unemployment mean that it is not always an easy matter for governments to devise appropriate policies to lower it.

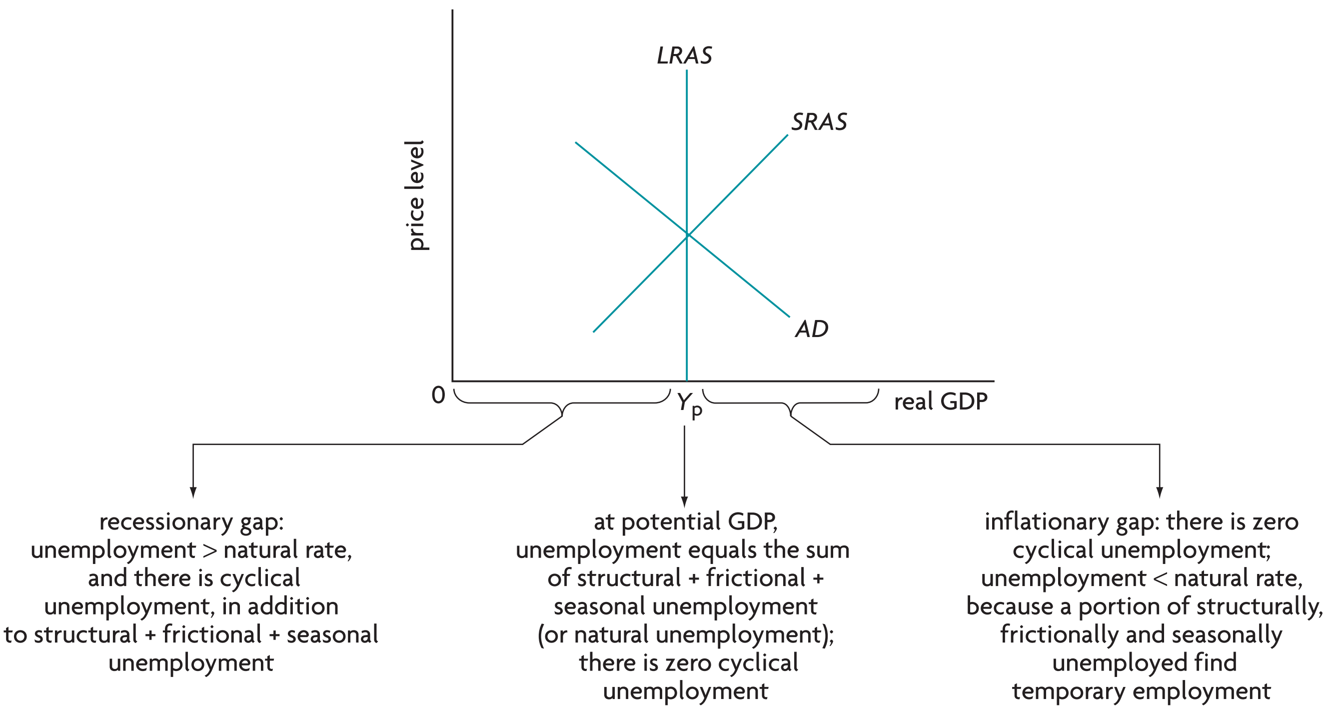


Figure 10.4: The four types of unemployment in relation to the *AD-AS* model

TEST YOUR UNDERSTANDING 10.2

1. Explain whether or not full employment mean the absence of unemployment.
2. Describe, using examples, the meaning and causes of structural, frictional and seasonal unemployment.
3. Use diagrams to show structural unemployment arising from
   a. a change in demand for labour skills,
   b. a change in the geographical location of industries, and
   c. higher than equilibrium wages.
4. Outline how the sum of structural, frictional and seasonal unemployment relates to the concepts of 'full employment' and potential output in the context of the *AD-AS* model.
5. Using a diagram, explain cyclical (demand-deficient) unemployment and the circumstances under which it arises.
6. Identify and outline the different kinds of unemployment an economy is likely to be experiencing when it is in
   a. a deflationary/recessionary gap,
   b. an inflationary gap, and
   c. when it is producing real GDP equal to potential GDP.

THEORY OF KNOWLEDGE 10.1

What is 'natural' about the natural rate of unemployment?

We have defined the natural rate of unemployment to be the sum of structural, frictional and seasonal unemployment, or more simply all unemployment other than that caused by the business cycle. To call this unemployment 'natural' appears strange; it suggests there is something normal, usual or standard about these types of unemployment. Yet, what can be 'natural' about particular types of unemployment?

It is also strange that natural unemployment corresponds to 'full employment'. Even though, as we know, it is never possible to have full employment in the sense of zero unemployment, why should the presence of structural, frictional and seasonal unemployment be known as 'full employment'?

We can understand the reasons behind the use of these terms if we put the natural rate concept in historical perspective. This concept was developed in the late 1960s by Milton Friedman, the founder of 'monetarism', for which he received the Nobel Prize. (Edmund Phelps, another Nobel Prize winner, also independently developed this concept.)

The natural rate concept is a reflection of the monetarist model, which views the economy as a stable system that automatically tends toward long-run equilibrium where there are no recessionary or inflationary gaps (see Chapter 9, Section 9.3). In this view, the forces of supply and demand work to allocate resources efficiently, and the most efficient allocation that is possible is reached when the economy is at long-run equilibrium. If the labour market worked completely perfectly according to supply and demand, unemployment would in fact drop to zero. However, in the real world this does not occur, because of institutions that lead to labour market imperfections: there will always be some people who cannot instantly find jobs due to lack of information, but in addition, and most importantly, people cannot find jobs because of the presence of labour market rigidities, including minimum wages, labour union activities and lack of incentives to work.

CONTINUED

In fact, Friedman believed (and monetarist economists continue to believe) that the natural rate of unemployment is caused only by labour market rigidities. In the long run, labour market rigidities can explain *all types of unemployment*, including frictional, seasonal and all of structural unemployment discussed above (cyclical unemployment, you may remember, is zero when the economy is at long-run equilibrium). If people become unemployed because industries change or move, in the long run in a free competitive market they would all respond to market incentives and they would all find jobs.

Therefore, in this view, in the long run, all unemployment arises *naturally* from an imperfectly working labour market. Since the labour market institutions that lead to this natural unemployment are considered as given (fixed) in any economy, it is 'reasonable' to consider that the economy has *full employment* when all workers are employed *except those who are unemployed due to the labour market imperfections*.

It is worth quoting Nobel Prize-winning economist Robert Solow once again:

> *'Whether the social scientist wills it or knows it, perhaps even if he fights it, his choice of research problem, the questions he asks, the questions he doesn't ask, his analytical framework, the very words he uses, are all likely to be, in some measure, a reflection of his interests, ideologies and values.'*[4] *(his analytical framework, the very words he uses)*

Friedman was an enthusiastic believer in the powers of the free market to solve the major economic problems, and he strongly opposed government intervention. His political convictions and values undoubtedly influenced his theoretical orientations, as well as his choice of words. His ideas had an important impact on policy-making, especially during the 1970s and 1980s, and his influence can still be felt to the present day. He was highly influential in the development of supply-side policies (see Chapter 13).

Some economists criticise his natural rate concept (as well as all his other ideas), especially since the 1990s. James K. Galbraith, a Keynesian economist (and son of the famous economist John Kenneth Galbraith), wrote the following:

> *'Alas, the location of the natural rate is not actually observed. Worse, the damn thing will not sit still. It is not only invisible, it moves! . . . [Economists] obsessively estimate and re-estimate the location of the natural rate, in order to guide their policy judgements. Sadly, they have never yet been able to predict its location . . .'*[5]

In addition to being difficult to estimate, the natural rate of unemployment is also imprecise as to its meaning. In this text, we follow the practice of most textbooks in using the term 'natural rate of unemployment' to refer to the sum of structural, frictional and seasonal unemployment. This is a convenient way to make a distinction between cyclical and all other types of unemployment. This does not presuppose that it is easy or even possible to actually measure any type of unemployment individually. As noted in the text, in the real world it is very difficult to distinguish between the different kinds of unemployment.

We end with a question. Why do economists bother to measure the 'natural rate of unemployment' and why is this so important? We will return to this question in the Theory of knowledge 10.2 at the end of this chapter (at HL).

[4] Robert M. Solow (1996) 'Science and ideology in economics,' in D. M. Hausman, *The Philosophy of Economics*, Cambridge University Press.
[5] James K. Galbraith (1996) 'The surrender of economic policy,' in *The American Prospect*, March–April.

CONTINUED

Thinking points

- Do the terms 'natural' rate of unemployment and 'full employment' have a normative aspect? What do they suggest in terms of government policy action (or inaction) to reduce the rate of unemployment?
- Can the use of language to reflect an underlying political ideology interfere with the use of the scientific method?
- Do you agree with Friedman (and other economists) that all non-cyclical unemployment can be explained in terms of labour market institutions that create rigidities in the labour market?

10.2 Low and stable rate of inflation

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between inflation, disinflation and deflation (AO2)
- explain how the inflation rate is measured using a consumer price index (AO2)
- calculate a weighted price index using quantities purchased as weights (HL only) (AO4)
- calculate the rate of inflation (AO4)
- explain the limitations of the CPI in measuring inflation (AO2)
- explain the causes of inflation: demand-pull and cost-push (AO2)
- draw diagrams illustrating demand-pull and cost-push inflation (AO4)
- explain and discuss the costs of inflation including redistributive effects, uncertainty, saving, export competitiveness, economic growth, inefficiency in resource allocation (AO3)[6]
- explain the causes of deflation: decreases in *AD* and increases in *SRAS* (AO2)
- draw diagrams illustrating deflation (AO4)
- explain the costs of deflation including redistributive effects, uncertainty, deferred consumption, high levels of cyclical unemployment, bankruptcies, real value of debt, inefficiency in resource allocation, ineffectiveness of policy (AO2)

[6] This point appears as AO2 in the syllabus, however it becomes AO3 in the point entitled 'Relative costs of unemployment versus inflation' considered later in this chapter.

Inflation, disinflation and deflation

Inflation is defined as a sustained increase in the general price level. When we speak of the 'general price level' we refer to an average of prices of goods and services in the entire economy, not to the price of any one particular good or service. 'Sustained' means that the general price level must increase to a new level and not fall back again to its previous lower level. Further, an increase in the general price level does not necessarily mean that prices of all goods and services are increasing; prices of some goods and services may be constant or even falling, while others are increasing. The presence of inflation indicates that prices of goods and services are increasing *on average*.

Deflation is defined as a sustained decrease in the general price level. As in the case of inflation, deflation refers to an average of prices; it is likely to be uneven, with some prices constant or even increasing.

Inflation is far more common than deflation; in fact, since the 1930s (the period of the Great Depression), most economies around the world have been experiencing a rising price level, or inflation.

Students sometimes confuse the difference between changes in the price level and changes in the rate of inflation. In our discussions of the *AD-AS* model, we have frequently seen increases in the price level; these indicate inflation. A change in the rate of inflation, by contrast, refers to a change in how fast the price level is rising. If the price level increases by 5% in one year and then increases by 7% the next year, this represents *an increase in the rate of inflation*. If the price level increases by 10% in one year and by 7% the next year, this represents *a decrease in the rate of inflation*, and is called **disinflation**. Disinflation therefore occurs when inflation occurs at a lower rate.

You must also be careful not to confuse a fall in the rate of inflation, or disinflation, with a fall in the price level, or deflation. A fall in the rate of inflation, such as from 10% to 7%, means that the price level is increasing at a lower rate, hence is disinflation. A fall in the price level indicates that deflation is occurring. Deflation can be thought of as negative inflation. For example, deflation of −2% means that the price level is falling at the rate of 2%.

Measuring inflation and deflation

The consumer price index

Measures of inflation (and deflation) are obtained by use of price indices (*indices* is the plural of *index*). A price index is a measure of average prices in one period relative to average prices in a reference period called a base period. One of the most commonly used price indices to measure inflation is the consumer price index (CPI).

The consumer price index (CPI) is a measure of the cost of living, or the cost of goods and services purchased by the typical household in an economy. It is constructed by a statistical service in each country, which creates a hypothetical 'basket' containing thousands of goods and services that are consumed by the typical household in the course of a year. The value of this basket is calculated for a particular year (called a base year); this is done by multiplying price times quantity for each good and service in the basket, and adding up to obtain the total value of the basket. The value of *the same basket* of goods and services is then calculated for subsequent years. The result is a series of numbers that show the value of the *same basket* of goods and services for different years. The CPI is then constructed to show how the value of the basket changes from year to year by comparing its value with the base year.

Once the consumer price index is constructed, inflation and deflation can be expressed as a percentage change of the index from one year to the other, which is simply a measure of *the percentage change in the value of the basket from one year to another*. Since the value of the basket changes from one period to another because of changes in the prices of the goods in the basket, these percentage changes reflect changes in the average price level. A rising price index indicates inflation; a falling price index indicates deflation. CPIs and rates of change in the price level are also calculated on a monthly basis and a quarterly basis.

The consumer price index (CPI) is a measure of the cost of living for the typical household, and compares the value of a basket of goods and services in one year with the value of the same basket in a base year. Inflation (and deflation) are measured as a percentage change in the value of the basket from one year to another. A positive percentage change indicates inflation. A negative percentage change indicates deflation.

Constructing a weighted price index (HL only)

We will construct a consumer price index (CPI) for a simple economy where consumers typically consume three goods and services: burgers, movie tickets and haircuts, shown in column 1 of Table 10.1. Column 2 gives us the quantities of each that the typical household buys in a year; these are the *weights*. Note that a **weighted price index** is a price index that 'weights' the various goods and services according to their relative importance in consumer spending.[7] To construct a CPI, we follow these steps:

1. Decide which of the years will be the base year; we choose 2017.
2. Use the price of each good and service in the base year (2017), to calculate its base year value (multiply quantity in column 2 by 2017 prices in column 3); these values appear in column 4.
3. Add up all values in column 4 to get the total value of the basket in the base year; this is $756, appearing at the bottom of column 4.
4. Use the price of each good and service in 2018 to calculate its 2018 value (multiply the number of units in the basket (column 2) by 2018 prices (column 5)); then do the same using 2019 prices (column 7); the resulting values appear in column 6 for 2018 and column 8 for 2019.
5. Add up the values in column 6 to obtain the total value of the basket in 2018; this is $798 appearing at the bottom of column 6; do the same to find the value of the basket in 2019, which is $900, appearing at the bottom of column 8.

We now have all the information we need to construct our price index for 2017, 2018 and 2019. Note that:

price index for a specific year

$$= \frac{\text{value of basket in a specific year}}{\text{value of same basket in base year}} \times 100$$

Therefore, the price index numbers for 2017, 2018 and 2019 are:

$$\text{price index for 2017} = \frac{756}{756} \times 100 = 1.00 \times 100 = 100.0$$

$$\text{price index for 2018} = \frac{798}{756} \times 100 = 1.055 \times 100 = 105.5$$

$$\text{price index for 2019} = \frac{900}{756} \times 100 = 1.190 \times 100 = 119.0$$

Note that the price index for the base year is always equal to 100.

1 Good and services	2 Quantity (number of units) in basket (weights)	3 Prices of basket goods and services in base year (2017)	4 Value of basket goods and services in base year (2017)	5 Prices of basket goods and services in 2018	6 Value of basket goods and services in 2018	7 Prices of basket goods and services in 2019	8 Value of basket goods and services in 2019
Burgers	37	$3	$111	$4	$148	$5	$185
Movie tickets	25	$15	$375	$14	$350	$16	$400
Haircuts	15	$18	$270	$20	$300	$21	$315
Total value of basket			$756		$798		$900

Table 10.1: Constructing a hypothetical price index

[7] In the real world, the weights used to construct the CPI are based on the proportion of consumer spending on each good or service on average, rather than on the quantity of each good or service consumed.

To construct a weighted price index, (i) find the value of the basket in current prices for each year; (ii) divide the value of the basket for each year by the value of the basket in the base year and multiply by 100. This will give you the price index number for each year.

In the real world, calculations of price indices are complicated as they involve collecting price data on thousands of goods and services and carrying out all necessary computations. This is done by specialised statistical services in every country.

Using a weighted price index (the CPI) to calculate the rate of inflation

A price index can be used to calculate the rate of inflation. Suppose we are given the following price index (it is actually the one calculated above for HL). The third row also shows the value of the basket for each of the years.

Year	2017	2018	2019
Consumer price index	100.0	105.5	119.0
Value of basket	$756	$798	$900

The rate of inflation is the percentage change in the price index. It is also given by the percentage change in the value of the basket. (This follows from the fact that the price index is calculated from the values of the basket.) The percentage change in a variable A is calculated by the following:

$$\%\text{ change in } A = \frac{\text{final value of A} - \text{initial value of A}}{\text{initial value of A}} \times 100$$

(See 'Quantitative techniques' chapter in the digital coursebook for more information.)

To calculate the percentage change in the price level from 2017 to 2018, we have the percentage change in price index from 2017 to 2018:

$$= \frac{105.5 - 100.0}{100.0} \times 100 = 5.5\%$$

In fact, we did not need to do this calculation: we can simply read the inflation rate from the price index, since $105.5 - 100.0 = 5.5\%$.

When the price level is presented as a price index, the rate of inflation is equal to the index number of any year minus the index number of the base year (which is always 100).

Therefore, it follows that the rate of inflation in the period 2017–2019 is $119.0 - 100.0 = 19.0\%$. However, it is only possible to read off the rate of inflation from a price index in this simple way in those cases involving a percentage change in the price level *relative to the base year*, whose price index number is equal to 100.[8] In other cases, we must use the formula above to calculate the rate of inflation. For example, to find the rate of inflation in 2018–2019:

$$\%\text{change in price index in 2018–2019}$$
$$= \frac{119.0 - 105.5}{105.5} \times 100 = 12.8\,\%$$

We could have found the same rates of inflation by calculating the percentage changes in the value of the basket:

$$\%\text{ change in value of the basket in 2017–2018}$$
$$= \frac{798 - 756}{756} \times 100 = 5.5\%$$

$$\%\text{ change in value of the basket in 2018–2019}$$

$= \dfrac{900 - 798}{798} \times 100 = 12.8\%$. We can see that these two percentages are the same as those calculated by use of the consumer price index.

Note that a price index with increasing values over time (such as the example above) indicates inflation. Decreasing values over time indicate deflation. Also, note that the first year in a price index need not be the base year. For example, suppose we have the following price index:

2000	2001	2002	2003	2004
97.5	100.0	107.3	109.7	107.8

The base year is 2001, for which the price index is 100. This price index indicates that inflation has occurred in 2000–2001, 2001–2002, and 2002–2003, but *deflation* has occurred in 2003–2004.

[8] Note that we can only use this rule for years *after* the base year and not before. For example, if the CPI is 92 in 2007 and 100 in 2008, we cannot say there is an 8% increase in the price level in the period 2007–2008. To convince yourself, do the calculation. You will find that the rate of inflation is 8.7%.

Calculating real income (Supplementary material)

In Chapter 8, we learned how to calculate real GDP from nominal GDP using the GDP deflator. We can now use the CPI to calculate real income (of consumers, pensioners, or other social groups):

$$\text{real income} = \frac{\text{nominal income}}{\text{CPI}} \times 100$$

Clearly, if nominal income increases by the same percentage as the price level (measured by the CPI), real income remains unchanged. The CPI is, in fact, very useful for calculating adjustments that must be made to nominal income (of wage-earners, pensioners, etc.) in order for these groups to maintain a constant or increasing real income.

A word of caution

Since the CPI compares price levels based on goods and services in a specific basket, it only makes sense to calculate inflation rates from a price index constructed by use of the same basket. Further, it is not possible to make comparisons of price levels (i.e. calculate rates of inflation) across years by use of price indices that have a different base year, even if the basket of goods and services is the same. Therefore, for comparisons of index numbers to be meaningful, the index numbers must be calculated using the same base year, and for the same basket of goods and services.

Comparing the CPI with the GDP deflator (Supplementary material)

If you would like to learn about the differences between the CPI and the GDP deflator, another price index we discussed in Chapter 8, you may read about this in the digital coursebook.

TEST YOUR UNDERSTANDING 10.3

1 Describe the meaning of the consumer price index (CPI) and explain the purpose for which it is constructed.

2 a Distinguish between inflation, deflation and disinflation, and provide numerical examples illustrating each of these.

b Explain, using examples, the difference between an increase in the price level and an increase in the rate of inflation.

3 Consider the following price index with corresponding values of the basket for the period 1997–2001:

Year	2014	2015	2016	2017	2018
CPI	97	95	100	105	107

a Identify the base year.

b Calculate the rate of inflation in the periods 2016–2017, and 2016–2018 without using the percentage change formula.

c Calculate the rate of inflation/deflation in 2014–2015, 2015–2016, and 2017–2018 using the CPI.

d Identify the period of time when deflation occurred.

e Outline whether disinflation occurred at any time.

4 Outline why it is important to use 'weights' for the goods and services consumed by the typical household.

5 (HL only) Using the data from the table,

a construct a consumer price index using 2009 as the base year.

b Identify the weights you are using.

c Calculate the rates of inflation/deflation for the years 2015–2016, 2016–2017, 2017–2018.

d Identify the years when inflation/deflation/disinflation occurred, and explain your conclusions.

e Construct a new price index using 2017 as the base year.

f Calculate the rates of inflation/deflation for the same three-year periods as in question (c).

CONTINUED

g Compare the rates of inflation/deflation you found using the two price indices (they should be the same!).

h Explain whether or not it would it make sense to compare an index number from the first price index with an index number from the second price index.

Good/ service	Quantity in basket	Price per unit in 2015 (£)	Price per unit in 2016 (£)	Price per unit in 2017 (£)	Price per unit in 2018 (£)
Pizzas	25	7	6	7	6
Movie tickets	9	15	17	18	18
Bus rides	47	2	4	4	3

Problems with the consumer price index (CPI)

- **Different rates of inflation for different income earners.** The rate of inflation calculated by use of the CPI reflects the change in average prices of goods and services included in the basket. However, different consumers have different consumption patterns depending on their income levels, and these may differ from what is included in the basket. This means they face different rates of inflation than what is calculated on the basis of the CPI basket.
- **Different rates of inflation depending on regional or cultural factors.** Exactly the same idea as above applies to consumer groups whose purchases differ from the typical household's consumption patterns, because of variations in tastes due to cultural and regional factors.
- **Changes in consumption patterns due to consumer substitutions when relative prices change.** Each good and service included in the basket is weighted (multiplied by the number of units of the good or service purchased by the typical household over a year). However, as some goods and services become cheaper or more expensive over time, consumers make substitutions, buying more units of the cheaper goods and less of the more expensive ones. This results in changing weights, but because the weights in the basket are fixed, the changes in consumption patterns cannot be accounted for in the CPI. Therefore, the CPI gives a misleading impression of the degree of inflation, usually overstating it.
- **Changes in consumption patterns due to increasing use of discount stores and sales.** In many countries, consumers increasingly make use of discount stores and sales, thus buying some goods and services at lower prices than those used in CPI calculations. This is another reason why the CPI tends to overstate inflation.
- **Changes in consumption patterns due to introduction of new products.** In this case, too, a fixed basket of goods and services cannot account for new products introduced into the market, as well as older products that become less popular or are withdrawn (consider for example the replacement of DVDs by on-demand services such as Netflix).
- **Changes in product quality.** The CPI cannot account for quality changes over time.
- **International comparisons.** The CPIs of different countries differ from each other with respect to the types of goods and services included in the basket, the weights used and methods of calculation. This limits the comparability of CPIs and inflation rates from country to country. To address this problem, the European Union (EU) has devised a Harmonised Index of Consumer Prices (HICP). The HICP determines consistent and compatible rules that must be followed by EU countries in

order to calculate CPIs that are consistent with each other.[9]

- **Comparability over time.** Virtually all countries around the world periodically revise their CPI baskets and change the base year (usually about every ten years) to try to deal with many of the problems noted above. In many countries the weights of goods and services are changed as often as every year. This means that whereas price index numbers are comparable over short periods of time, over longer periods comparability is lessened because of cumulative changes in the basket of goods and services.

The core rate of inflation (Supplementary material)

There are certain goods, notably food and energy products (such as oil) that have highly volatile prices (meaning they fluctuate widely over short periods of time). Reasons for price volatility include wide swings in supply or demand, causing large and abrupt price changes. When such goods are included in the CPI, they may give rise to misleading impressions regarding the rate of inflation. To deal with this problem, economists measure a *core rate of inflation,* which usually is done by constructing a CPI that does not include food and energy products with highly volatile prices.

TEST YOUR UNDERSTANDING 10.4

1. Explain why the CPI may not be an accurate measure of the rate of inflation.
2. (Optional) Describe the meaning of a core rate of inflation and how this is calculated.

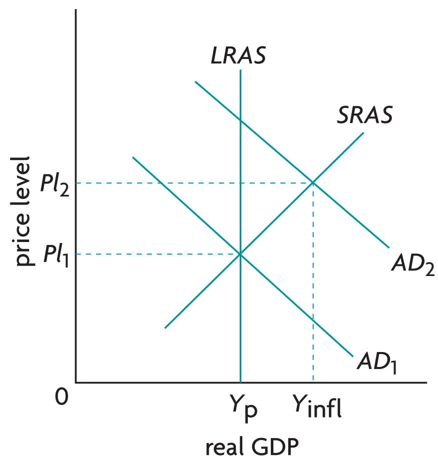
Causes of inflation

We will examine two causes of inflation: demand-pull inflation and cost-push inflation.

Demand-pull inflation

Demand-pull inflation is caused by increases in aggregate demand, in turn brought about by changes in any of the determinants of aggregate demand (see Chapter 9, Section 9.1). Assume the economy is initially at full employment equilibrium, producing potential GDP, shown as Y_p in Figure 10.5(a) and (b). The economy experiences an increase in aggregate demand appearing as a rightward shift of the AD curve from AD_1 to AD_2 in both diagrams. The impact on the economy is to increase the price level from Pl_1 to Pl_2, and to increase the equilibrium level of real GDP from Y_p to Y_{infl}. The increase in the price level from Pl_1 to Pl_2 due to the increase in aggregate demand is known as demand-pull inflation.

a The monetarist/new classical model



b The Keynesian model

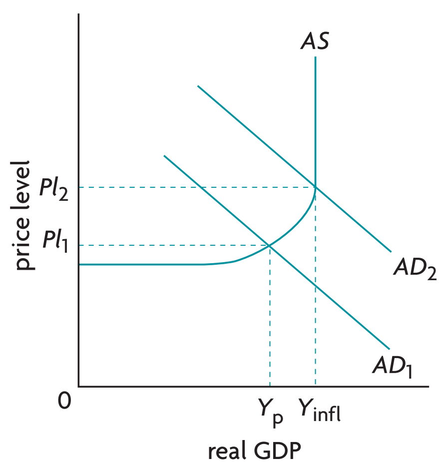


Figure 10.5: Demand-pull inflation

[9] The HICP does not determine a uniform basket for all countries (this is done in recognition of the point noted above that different regions/countries have different consumption patterns due to diverse tastes, cultural factors and income levels). The HICP succeeds to a large extent in resolving the comparability problem, and whereas it is not intended to replace national CPIs, it is used in all cases where comparisons across countries need to be made. The HICP is calculated by all European Union countries plus Iceland and Norway. The first base year to be used was 1996, and the index began being calculated from January 1997.

Note that demand-pull inflation is associated with an inflationary gap: real GDP is greater than full employment GDP, and unemployment falls to a level below the natural rate of unemployment. The demand for labour is so large that some workers who are structurally, frictionally or seasonally unemployed temporarily find jobs.

> **Demand-pull inflation** involves an excess of *aggregate demand* over *aggregate supply* at the full employment level of output, and is caused by an increase in aggregate demand. It is shown in the *AD-AS* model as a rightward shift in the *AD* curve.

Cost-push inflation

Cost-push inflation is caused by increases in costs of production or supply-side shocks. Assume the economy is initially at the full employment level of output, Y_p in Figure 10.6, and suppose there is an increase in costs of production. The *SRAS* curve shifts from $SRAS_1$ to $SRAS_2$, leading to an increase in the price level from Pl_1 to Pl_2, and a fall in the equilibrium level of real GDP from Y_p to Y_{rec}. The increase in the price level due to the fall in *SRAS* is known as cost-push inflation.

Cost-push inflation is analysed only by means of the monetarist/new classical *AD-AS* model. The Keynesian model is not equipped to deal with short-term fluctuations of aggregate supply. Keynes was concerned with showing the importance of *aggregate demand* in causing short-term fluctuations. The output level Y_{rec}, though indicating a recession, is not called a recessionary/deflationary gap, because output gaps

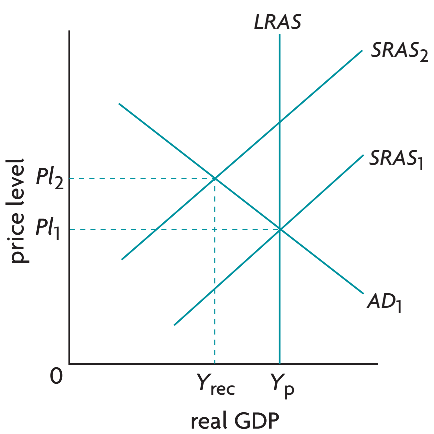


Figure 10.6: Cost-push inflation

(whether recessionary or inflationary) can only be caused by too little or too much aggregate demand (see Chapter 9, footnote 4).

In Chapter 9, we saw that a decrease in *SRAS* poses a special set of problems because it leads to both inflation and a fall in real GDP (with more unemployment; see Section 9.3). The presence of both inflation and unemployment is called *stagflation*, a combination of the words 'stagnation' and 'inflation'. This should be contrasted with an increase in aggregate demand leading to demand-pull inflation, which results in a higher price level but an *increase* in real GDP (with less unemployment). Cost-push inflation, or stagflation, is more difficult to deal with effectively, as we will discover in Chapter 13.

> **Cost-push inflation** is caused by a fall in aggregate supply, in turn resulting from increases in wages or prices of other inputs, shown in the *AD-AS* model as leftward shifts of the *AS* curve.

TEST YOUR UNDERSTANDING 10.5

1. Using appropriate diagrams, explain the difference between demand-pull and cost-push inflation.
2. Outline why cost-push inflation is potentially more serious than demand-pull inflation,
3. Using diagrams, show the effects on the price level of the following events, and explain whether it is cost-push or demand-pull inflation.
   a. Real GDP in foreign countries that trade with your country increases, leading to increased demand for your country's exports.
   b. Businesses are optimistic that a recession is about to end, and so increase investment spending.
   c. An increase in housing prices makes consumers increase their consumption expenditures.
   d. A sudden increase in the price of oil, a key input in production, occurs.

REAL WORLD FOCUS 10.2

Rising inflation in Pakistan

The consumer price index in Pakistan has nearly tripled in the course of one year, rising from 3.2% in March 2018 to 9.4% a year later, reaching the highest inflation rate in seven years. This has been the result of a mix of demand-pull and cost-push factors. The core rate of inflation, which excludes food and energy prices, is only slightly lower at 8.5%. Increases in energy prices and rising costs of production, have been responsible for cost-push inflation. Demand-pull inflation has been the result of increasing demand, itself the result of an increase in the supply of money. It is expected that inflation will soon rise further to double digits.

Source: *https://fp.brecorder.com*

Figure 10.7: Karachi, Pakistan. Empress Market Building, surrounded by numerous markets and shops

Applying your skills

1 Explain the meaning and use of the
   a consumer price index (CPI),
   b core inflation rate.

2 Using diagrams, explain what factors have resulted in cost-push and demand-pull inflation in Pakistan.

Costs of a high rate of inflation

Inflation, and especially a high rate of inflation, poses problems for an economy, because it affects particular population groups especially strongly, as well as the economy as a whole.

The relationship between inflation, purchasing power and nominal and real income

To understand why problems can arise, let's consider the relationship between inflation and purchasing power, and nominal and real income. Purchasing power refers to the quantity of goods and services that can be bought with money. Imagine you have £60 to spend on shirts. You can think of this as your 'nominal income'. When the price is £20 per shirt, you can buy three shirts. If the price increases to £30 per shirt, you can only buy two shirts. Your money, or your nominal income of £60 has not changed, yet the purchasing power of the £60, or what this money can buy, has fallen due to the increase in price. 'Real income' is the same as 'purchasing power'; it refers to what your money can buy: it decreases as prices rise, and increases as prices fall.

Changes in real income, money income and the general price level are related to each other in the following way:

% change in real income (or purchasing power) = % change in nominal income − % change in the price level (or the rate of inflation)

These relationships illustrate some important points. Inflation leads to a fall in real income, or purchasing power, only if nominal income is constant, or if nominal income increases more slowly than the price level. Say there is a 5% increase in the price level, which is a 5% rate of inflation. How will your real income be affected? If your nominal income also increases by 5%, your real income, or purchasing power, remains unchanged. Therefore, for you, inflation is not a problem. If, however, your nominal income remains constant or increases by less than 5%, your real income falls, and you will be worse off since the purchasing power of your income is reduced.

Costs of inflation

Redistribution effects

Inflation redistributes income away from certain groups in the economy and towards other groups. Redistribution arises in situations where certain groups lose some purchasing power and become worse off, while other groups gain purchasing power and become better off. Groups who lose from inflation include:

- **People who receive fixed incomes or wages.** When individuals receive an income or wage that is fixed or constant, as the general price level increases they become worse off. This occurs when:
  - workers have wage contracts fixing their wages over a period of time
  - pensioners receive fixed pensions
  - landlords receive fixed rental income
  - individuals receive fixed welfare payments.
- **People who receive incomes or wages that increase less rapidly than the rate of inflation.** When individuals' incomes do not keep up with a rising price level (do not increase as fast as the price level), a fall in their real incomes results and they therefore become worse off. These groups may include all those noted above plus any other kind of income receiver whose income is not increasing as rapidly as the price level.
- **Holders of cash.** As the price level increases, the real value or purchasing power of any cash held falls.
- **Savers.** People who save money may become worse off as a result of inflation. In order to maintain the real value of their savings, savers must receive a rate of interest that is at least equal to the rate of inflation. Suppose you deposit $1000 in a bank account that pays you no interest. If there is inflation, the real value of your savings will fall. However, you may be able to protect the purchasing power of your savings. Say the rate of inflation is 5% per year. If you receive interest on your deposit at the rate of 5% per year, what you will lose through inflation will be exactly matched by what you gain through interest income. In this case, the real value (or purchasing power) of your savings remains unaffected. In general, savers who receive a rate of interest on their savings lower than the rate of inflation suffer a fall in the real value (or purchasing power) of their savings.
- **Lenders (creditors).** People (or financial institutions such as banks) who lend money may be worse off due to inflation. Assume you lend your friend €100 for one year (and you do not charge interest). If in the course of the year there is an increase in the price level (inflation), the real value of the €100 you will get back from your friend at the end of the year will have fallen. If you charged your friend a rate of interest equal to the rate of inflation, then the real value of your loan to your friend will be exactly maintained. In general, lending at a lower interest rate than the rate of inflation makes the lender (creditor) worse off at the end of the loan period.

Groups who gain from inflation include:

- **Borrowers (debtors).** In the example above, your friend who borrowed €100 from you benefits since the €100 paid back after one year is worth less than one year ago. If you had charged interest, your friend (the borrower) would benefit as long as the rate of interest is lower than the rate of inflation. In general, borrowing at a lower interest rate than the rate of inflation makes the borrower (debtor) better off at the end of the loan period.
- **Payers of fixed incomes or wages.** As long as nominal wages, pensions, rents, welfare payments, etc., are fixed while there is inflation, the payers (whether they are firms, the government, payers of rent, etc.) benefit as the real value of their payments falls due to inflation.
- **Payers of incomes or wages that increase less rapidly than the rate of inflation.** As long as incomes of any kind increase less rapidly than the rate of inflation, the payers of these incomes benefit due to the falling real value of their payments.

Uncertainty

Inability to accurately predict what inflation will be in the future means that people cannot predict future changes in purchasing power (of income, wealth, loans and anything else that is measured in terms of money). This causes uncertainty among economic decision-makers. Firms, in particular, become more cautious about making future plans under uncertainty about future price levels, because they are unable to make accurate forecasts of costs and revenues. Their uncertainty leads them to make fewer investments, which may lead to lower economic growth.

Effects on saving

We saw above that when there is inflation, savers lose if they receive no interest on their savings or if the rate of interest on their savings is lower than the rate of inflation. Therefore, inflation lowers the incentive to save. Further, if the rate of inflation is high, people may spend more now in order to avoid higher prices in the future, in which case the effect may be to further lower saving.

International (export) competitiveness

When the price level in a country increases more rapidly than the price level in other countries with which it trades, its exports become more expensive to foreign buyers, while imports become cheaper to domestic buyers. The country's international competitiveness, or its ability to compete with foreign countries, is reduced. The result is that the quantity of exports falls, and the quantity of imports increases. This in turn may create difficulties for the country's balance of payments (see Chapter 16).

Effects on economic growth

High inflation does not favour economic growth. As we have seen above, among the consequences of inflation are uncertainty among firms, which causes investment to fall; in addition lower saving means that there are less funds available for investment. These two factors lead to drops in investment, which is a component of aggregate demand, therefore to a fall in aggregate demand. Moreover, we have also seen that inflation leads to lower exports and higher imports, both contributing to a fall in net exports, which also causes a fall in aggregate demand. Falling aggregate demand leads to lower real GDP.

Effects on resource allocation

As we discussed in Chapter 2, the price mechanism plays an important role in resource allocation. If prices are rising rapidly, the signalling and incentive functions do not work effectively. The reason is that prices do not increase in the same proportion for all products, they rise more for some products than for others, meaning that the signals and incentives they provide for consumers and producers become distorted and therefore inaccurate. The result is that allocative inefficiency is increased.

Social and personal costs that are unequally distributed

In view of the redistribution effects of inflation, we have seen that people on fixed incomes suffer losses as their income loses its purchasing power. This often includes pensioners, unemployed people receiving unemployment benefits, and also workers whose wages are either fixed or do not rise as fast as the rate of inflation. In addition, people on low incomes are not in position to place their savings in assets that do not lose their value with inflation, such as real estate, stocks in the stock market, gold or even jewellery. Further, rising prices of necessities such as food and energy needed for heating can cut deeply into the incomes of lower income people. Therefore it is likely that people on low incomes are more seriously affected by high rates of inflation than people on higher incomes.

Consequences of hyperinflation

Hyperinflation consists of very high rates of inflation. It is defined as occurring when the price level increases by more than 50% per month, though it can reach thousands or even millions of percentage points per year. One of the most dramatic hyperinflations in history occurred in Germany after the First World War, when the price level in 1924 was more than 100 trillion times higher than in 1914. In more recent years, many hyperinflations have been concentrated in Latin America from the mid-1980s to the early 1990s, and in eastern European and former Soviet Union countries in the early 1990s following the collapse of the Soviet Union. Peak annual rates of inflation came to about 7500% in Peru in 1990; 3080% in Argentina in 1989; 2950% in Brazil in 1990; 1735% in Russia in 1992; 4735% in Ukraine in 1995; and 1060% in Bulgaria in 1997. One of the most serious cases of hyperinflation occurred in Zimbabwe, where the rate of inflation went from over 1000% in 2006, to 12 000% in 2007, and to over 11 million % (on an annual basis) in the summer of 2008. In Venezuela in 2018 inflation was an estimated 80 000%.

Hyperinflation results from very significant increases in the supply of money, which impact directly on the price level. Hyperinflations occur when governments resort to printing money, thereby increasing its supply.

Hyperinflation has serious negative consequences, over and above those discussed above, because money loses its value very rapidly. Consumers increase their spending to benefit from the current prices before they increase in the future, thereby feeding aggregate demand, which causes demand-pull inflation. Workers demand higher nominal wages to maintain the real value of their current and future incomes, thereby feeding cost-push inflation. Therefore, an *inflationary spiral* is created (a process where inflation sets in motion a series of events that worsen the inflation).

Serious hyperinflations result in a massive disruption of economic activity: businesses stop investing in productive activities and invest instead in assets that are believed to maintain their value as prices rise (gold, real estate or jewels); firms also withhold goods from sale in the market so that they can sell them later at higher prices; lenders (creditors) suffer massive losses as the real value of debts falls dramatically. At the extreme, money loses its value altogether and people resort to barter (the direct exchange of goods or services, eliminating the need for money), which in itself makes production and exchange extremely difficult. Serious hyperinflations can also lead to political and social unrest.

What is an appropriate rate of inflation?

Most governments prefer a *low and stable rate of inflation*, not a zero rate of inflation. The reason why a zero rate of inflation, meaning a constant price level, is not preferred is that this comes dangerously close to deflation, which as we will see below can cause serious problems for an economy.

There is no one particular rate of inflation that is ideal, but many governments would like to see this in the range of about 2–3% per year. Less than 2% might be considered as coming close to deflation; more than 4% is seen as being too high.

TEST YOUR UNDERSTANDING 10.6

1. Using a numerical example, explain the relationship between inflation and purchasing power.
2. Explain what happens to your real income (your purchasing power) in each of the following situations:
   - a your nominal income increases by 5% and the rate of inflation is 8%,
   - b your nominal income falls by 10%, and the rate of inflation is 3%, and
   - c your nominal income increases by 7% and the rate of inflation is 7%.
3. a Inflation results in redistribution of purchasing power. Explain who is likely to gain and who is likely to lose from the redistribution effects of inflation.
   b Identify and explain some other negative consequences of inflation (other than redistribution).

Causes of deflation

Why deflation occurs rarely in the real world

Deflation, a falling price level on average, is not a common phenomenon. Whereas it is often the case that the price of a particular good or service may fall over time, it is rare to see the general price level of an economy falling. There are several factors that account for this:

- **Wages of workers do not ordinarily fall.** This means it is difficult for firms to lower the prices of their products, as this would cut into their profits, especially since wages represent a large proportion of firms' costs of production. There are several reasons why wages do not fall easily (labour contracts, minimum wage legislation, worker and union resistance to wage cuts, ideas of fairness, fears of negative impacts on workers' morale, etc.)
- **Large oligopolistic firms may fear price wars.** If one firm lowers its price, then others may lower theirs more aggressively in an effort to capture market share, and then all the firms will be worse off. Therefore, firms avoid cutting their prices.

Whereas deflation occurs rarely, it has appeared periodically, for example in Britain and the United States in the late 19th century, in the United States during the Great Depression of the 1930s (1933–1937), and in Japan from 1999 to 2006. In 2003 and again in 2008, there were serious concerns in Europe and the United States that deflation might occur. Deflation is generally feared more than inflation for reasons we will discover below.

Causes of deflation

We can make a distinction between two causes of deflation: decreases in aggregate demand and increases in aggregate supply. These can be seen in Figure 10.8.

In Figure 10.8(a) a decrease in aggregate demand (for example, due to business pessimism) causes the AD curve to shift from AD_1 to AD_2. Whereas the AD-AS model predicts a drop in the price level, or deflation, this is unlikely to occur over a short period of time for the reasons discussed above, accounting for the highly infrequent occurrence of deflation.[10] However, if low aggregate demand persists over a long period, the

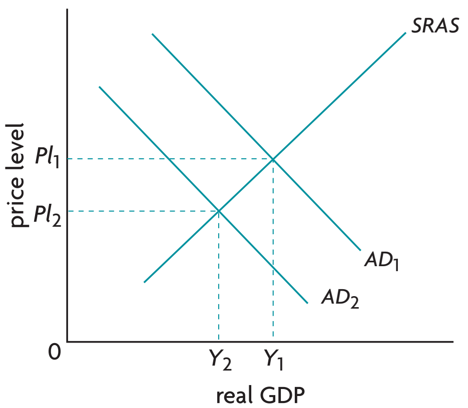
[10] The infrequent occurrence of deflation will be explained by the 'ratchet effect', to be discussed in Chapter 13.

price level falls to Pl_2. This is sometime referred to as 'bad' deflation because it is associated with recession, falling incomes and output since real GDP falls from Y_1 to Y_2, and cyclical unemployment. These are the circumstances that characterised the deflation of the Great Depression during the 1930s and, more recently, in Japan.

Figure 10.8(b) shows a rightward shift of the *SRAS* curve, with the *AD* curve constant, which gives rise to a new point of equilibrium that occurs at a lower price level. This is sometimes referred to as 'good' deflation because it is associated with economic expansion since real GDP increases from Y_1 to Y_2, rising incomes and output, increasing employment and economic growth. Some economists argue that it was under such circumstances that the deflation of Britain and the United States in the late 19th century occurred.

However, it must be stressed that while it may be possible to make a theoretical distinction between 'good' and 'bad' deflation, *no deflation is ever good*. We will discover the reasons for this in the following pages, where we will see that deflation discourages spending, causing aggregate demand to fall regardless of the causes of deflation. Deflation is therefore considered by economists to be a greater threat than inflation.

a Falling aggregate demand (*AD*)



b Increasing short-run aggregate supply (*SRAS*)

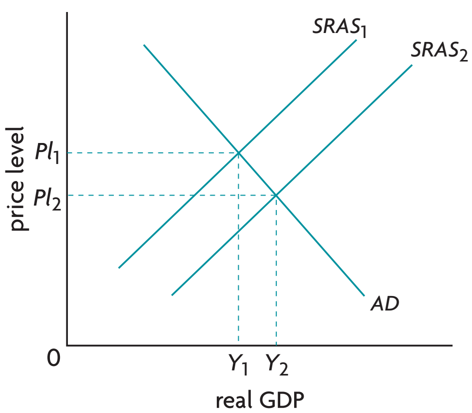


Figure 10.8: Causes of deflation

Costs of deflation

Redistribution effects

The redistribution effects of deflation are the opposite of those of inflation: with a falling price level, individuals on fixed incomes, holders of cash, savers and lenders (creditors) all gain as the real value of their income or holdings increases. By contrast, borrowers (debtors) and payers of individuals with fixed incomes lose with a falling price level, as they must pay out sums that have an increasing real value.

Increase in the real value of debt

In view of the above, the real value of debt increases. If you hold $1000 and the price level falls, this means that the purchasing power of your money increases because you can buy more things with that amount. In just the same way if you owe this $1000, its real value in terms of its purchasing power increases when the price level falls.

Uncertainty

Deflation, like inflation, creates uncertainty for firms, which are unable to forecast their costs and revenues due to declining price levels.

Deferred consumption, high and increasing cyclical unemployment: risk of a deflationary spiral

Deferred consumption means that consumers postpone spending. Consumers postpone making purchases when they see falling prices as they expect that prices will continue to fall. Therefore, deflation discourages spending. Deflation also discourages borrowing by both consumers and firms, because the real value of debt increases as the price level falls. The result is that consumer and business spending falls, causing aggregate demand to fall. Falling *AD* results in lower real GDP with cyclical unemployment, and also causes the price level to fall further. This in turn gives rise to further postponement of spending, *AD* falls further, unemployment increases further, incomes and prices fall further, deflationary pressures increase further, and so on in a downward spiral. This is known as a deflationary spiral, shown in Figure 10.9.

Risk of bankruptcies and a financial crisis

As we saw above, deflation results in an increase in the real value of debt. If the economy is in recession, and incomes

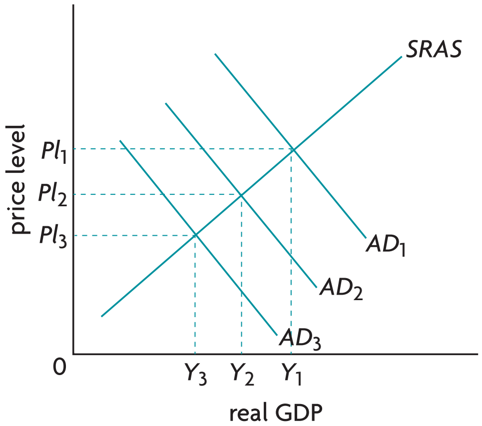


Figure 10.9: Deflationary spiral

are falling while the real value of debt is increasing, the result will most likely be bankruptcies of firms and consumers who are unable to pay back their debts. If such bankruptcies become widespread, banks and financial institutions will be affected, and a large risk of a major financial crisis arises.

Inefficient resource allocation

As we saw earlier, high rates of inflation lead to inefficient resource allocation because the signalling and incentive functions of prices are unable to work effectively. In deflation, prices of all goods and services do not fall uniformly, with the result that the price signals and incentives get distorted, leading to resource misallocation.

Policy ineffectiveness

Once a deflation sets in, it may be difficult for policy makers to deal with it. One reason is that when people's expectations of a falling price level become well established, and they get used to spending less in expectation of falling prices, it may be difficult for them to change their mindset. Another very important reason is that expansionary monetary policy (to be discussed in Chapter 13) may become ineffective. This monetary policy involves decreases in the rate of interest in order to encourage more borrowing and spending by consumers and firms, however once interest rates approach zero they cannot continue to fall. Therefore, monetary policy cannot be relied upon to solve the problem.

Summing up deflation

High cyclical unemployment, together with the risks of a deflationary spiral and a financial crisis, reveal the special and potentially serious dangers of deflation. These are worsened by the difficulties of finding solutions to the problem of deflation.

You can now see why governments prefer a *low and stable rate of inflation*, with a target rate of around 2% per year, which is sufficiently above zero so as not to be too close to deflation.

In addition, you can see why the distinction between 'good' and 'bad' deflation is actually meaningless, because once deflation sets in it does not matter how it originated since a deflationary spiral may result even from 'good' deflation.

A positive effect of deflation

In spite of all the negatives of deflation, there is one positive effect worth mentioning. While inflation hurts international competitiveness, deflation brings benefits. A lower price level means that exports may increase as foreigners now prefer buying from the country with lower prices while imports may fall as domestic consumers prefer the lower-price domestically produced goods. Therefore net exports $(X - M)$ increase, putting an upward pressure on aggregate demand and real GDP. However, this positive effect is not enough to counteract all the risks and dangers of deflation.

TEST YOUR UNDERSTANDING 10.7

1. Outline why deflation occurs rarely in the real world.
2. Explain what happens to your real income (your purchasing power) in each of the following situations:
   a. your nominal income increases by 5% and the rate of deflation is 3%,
   b. your nominal income falls by 10%, and the rate of deflation is 2%, and
   c. your nominal income falls by 3% and the rate of deflation is 4%. (You should think of deflation as negative inflation.)
3. Using diagrams, explain two possible causes of deflation.
4. a. Explain some negative consequences of deflation.

CONTINUED

b Explain why deflation may be especially serious for an economy.

5 Outline why a low, but above zero, rate of inflation is desirable.

10.3 Exploring the relationship between unemployment and inflation

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- discuss the relative costs of unemployment versus inflation (AO3)
- discuss the potential conflict between low unemployment and low inflation (AO2)[11]
- discuss the trade-off between unemployment and inflation: the short-run and long-run Phillips curves (HL only) (AO3)
- draw diagrams using the *AD-AS* model and showing the short-run and long-run Phillips curves (HL only) (AO4)

The relative costs of unemployment versus inflation

In Sections 10.1 and 10.2 above, we discovered that both unemployment and inflation have a number of costs. You should be able to identify what these are and discuss them.

During the 1970s, a well-known U.S. economist, Arthur Okun, created a 'misery index' which consists of the sum of the unemployment rate and the inflation rate of a country. The higher the index, the greater the misery of a population. However, the misery index does not distinguish between the separate effects of unemployment and inflation on the well-being, or lack of well-being, of a population. Can we say anything about which of the two contributes more to misery, or has greater costs? From a purely economic perspective, it is difficult to generalise about which of the two is less desirable, as this is likely to depend on the particular circumstances of the economy being considered, as well as on how high is the unemployment versus how high is the inflation.

In addition to the economic costs, we have seen that both unemployment and inflation have personal and social costs, over and above the economic ones. A number of studies that have examined the effects of unemployment versus inflation on well-being conclude that, between the two, unemployment has a stronger negative impact.

According to one such study that measured the loss of well-being, it was found that both unemployment and inflation increase unhappiness.[12] The study was based on a very large European dataset for the period 1975–2013, which included periods of high inflation as well as periods of high unemployment. It was found that unemployment increases unhappiness far more than inflation. Specifically *an increase of one percentage point in unemployment lowers well-being nearly six times more than a one percentage point increase in inflation*. Unemployment lowers the happiness of not only the unemployed but also the people around them, with women and the elderly being relatively more affected.

This is perhaps hardly surprising as the loss of a job leading to unemployment and therefore loss of income has potentially very serious financial, personal and social effects on the unemployed individuals and their families.

[11] Note that this bullet point as well as the next two come from the syllabus section 'Potential conflict between macroeconomic objectives'.
[12] https://www.dartmouth.edu/~blnchflr/papers/BLANCHFLOWER_et_al-2014-Journal_of_Money%252c_Credit_and_Banking.pdf

The conflict between low unemployment and low inflation

There are a number of potential conflicts between macroeconomic objectives, which we will study in Chapter 11. One of these conflicts is between the objectives of low unemployment and low inflation. This can be understood by use of the Keynesian *AD-AS* model, shown again in Figure 10.10. When there is a deflationary gap and the economy is in recession such as at output level Y_1 or Y_2, the rate of inflation is low (the price level is constant) but there is high cyclical unemployment. As aggregate demand increases and the economy approaches potential output, Y_p, the price level begins to rise while cyclical unemployment falls. As aggregate demand continues to increase, the price level increases even faster, while unemployment continues to fall. If aggregate demand increases further resulting in an inflationary gap such as at Y_4, unemployment falls to a level that is even lower than the natural rate, since some of the structurally, seasonally and frictionally unemployed will now find employment (see Figure 10.4 for an explanation).

The reason behind the increasing inflationary pressures is that as aggregate demand increases, resources are used more fully, giving rise to bottlenecks that result in higher wages and other resource prices. This process gives rise to higher product prices and hence a rising price level.

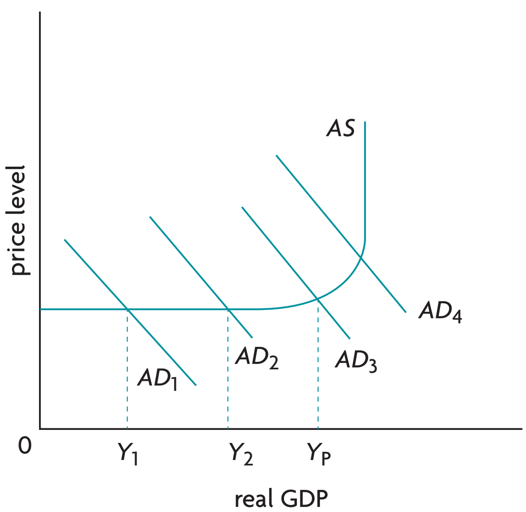


Figure 10.10: Keynesian model showing increasing price level with decreasing unemployment

It can therefore be concluded that it may be difficult to achieve both a low rate of inflation and a low unemployment rate at the same time.

The trade-off between unemployment and inflation (HL only)

The short-run Phillips curve

The **Phillips curve** is concerned with the relationship between unemployment and inflation. In the late 1950s, a New Zealand economist A.W. Phillips published a study showing that there appeared to be a long-term negative relationship between the unemployment rate and the rate of change in nominal (money) wages; this relationship was later extended by economists to apply to the relationship between unemployment and inflation. The relationship showed that the lower the rate of inflation, the higher the unemployment rate; and the higher the rate of inflation, the lower the unemployment rate. This relationship is shown in Figure 10.11(a), where the unemployment rate is measured along the horizontal axis, and the rate of inflation along the vertical axis. (Note that the vertical axis does not measure the price level, as in the *AD-AS* model.)

The Phillips curve suggests that if there is a constant negative relationship between the two variables, then every economy faces a trade-off between inflation and unemployment: it can choose between a relatively low rate of inflation and a higher unemployment rate, such as point a on the curve, or a higher rate of inflation and a lower unemployment rate, such as point d. Whereas, ideally, it would be preferable for any economy to have low inflation and low unemployment, such as point e, this is not possible according to the theory of the Phillips curve, as the only achievable points are those on (or close to) the curve.

The reasoning behind the shape of the curve can be illustrated by use of the *AD-AS* model, shown in Figure 10.11(b). Assume a fixed, upward-sloping *SRAS* curve, and imagine a succession of aggregate demand increases (which could be caused by any of the factors we are familiar with from Chapter 9 (Table 9.1)). As aggregate demand shifts from AD_1 to AD_2, the price level rises from Pl_1 to Pl_2, the level of real GDP increases from Y_1 to Y_2, and the level of unemployment correspondingly falls. The same process is repeated as aggregate demand increases from AD_2 to AD_3, and then to AD_4, and so on.

With every increase in aggregate demand, we have an increase in the price level and a fall in unemployment. It follows, then, that we can simply think of each point on the Phillips curve (such as a, b, c or d) as corresponding to the point of intersection of *SRAS* with a different *AD* curve (a, b, c or d). The 'choice' of where to be on the Phillips curve in part (a) thus corresponds to a 'choice' of *AD* curve in part (b) of the figure.[13]

a The shape of the Phillips curve

b The reasoning behind the Phillips curve in terms of the *AD-AS* model

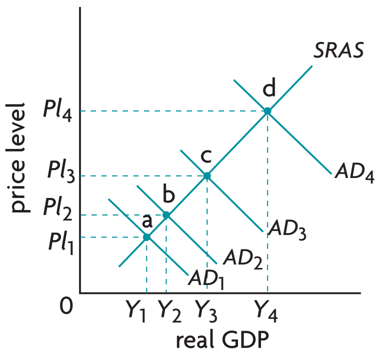


Figure 10.11: The short-run Phillips curve

The breakdown in the relationship: stagflation

During the 1960s, many economists came to believe that the Phillips curve did offer the possibility of choice between inflation and unemployment. At that time aggregate supply was relatively stable, and major changes in economic activity were caused by swings in aggregate demand. Most economists at the time were very strongly influenced by Keynesian thinking, believing that demand-side policies (see Chapter 13) were very important in influencing the level of economic activity and real GDP. The Phillips curve appeared to offer governments the possibility of using demand-side policies to choose between various alternatives. High aggregate demand would lead to low unemployment and higher inflation, while low aggregate demand would lead to higher unemployment and lower inflation.

Events of the 1970s and 1980s upset this line of thinking, and the stable relationship between inflation and unemployment that was suggested by the Phillips curve appeared to break down. Whereas it had been supposed that aggregate supply could remain stable over long periods of time, a number of aggregate supply shocks led to a period of *stagflation*, a term coined at the time to refer to the new phenomenon of stagnation (or recession) with unemployment and inflation simultaneously. The most important of the supply shocks involved the oil price increases brought on by the actions of OPEC (Organization of the Petroleum Exporting Countries), which restricted the global supply of oil. Another supply shock involved food price increases resulting from worldwide crop failures, restricting the global supply of food.

The impacts of these events on the Phillips curve and on the *SRAS* curve are shown in Figures 10.12(a) and (b). In part (b), we see that as the supply shocks cause the *SRAS* curve to shift leftward from $SRAS_1$ to $SRAS_2$ and then to $SRAS_3$, the result is higher price levels (from Pl_1 to Pl_2 and Pl_3) and lower levels of GDP (from Y_1 to Y_2 and Y_3), signifying increases in unemployment. In other words, decreases in *SRAS* (with *AD* constant) result in higher price levels and higher unemployment. This phenomenon is inconsistent with the logic of the Phillips curve, and was interpreted to involve *outward shifts in the Phillips curve*, which until then was thought to be stable and constant. The outward Phillips curve shifts appear in part (a), indicating that higher rates of inflation are associated with higher rates of unemployment; the move from point a to b and c in part (a) correspond to points a, b and c in part (b).

[13] The correspondence between Figure 10.11(a) and (b) is not entirely accurate. The vertical axis of part (a) measures the rate of inflation, or the percentage increase in the price level. The vertical axis of part (b) measures the price level, which is very different from the rate of inflation. There can be increases in the price level with no increase in the rate of inflation and even with a decrease in the rate of inflation (or disinflation; for example, the rate of inflation increases by 5% in 2000 and by 3% in 2001). The succession of *AD* curves in part (b), leading to increasingly larger rises in the price level, has been drawn with this point in mind, even though it is still not accurate.

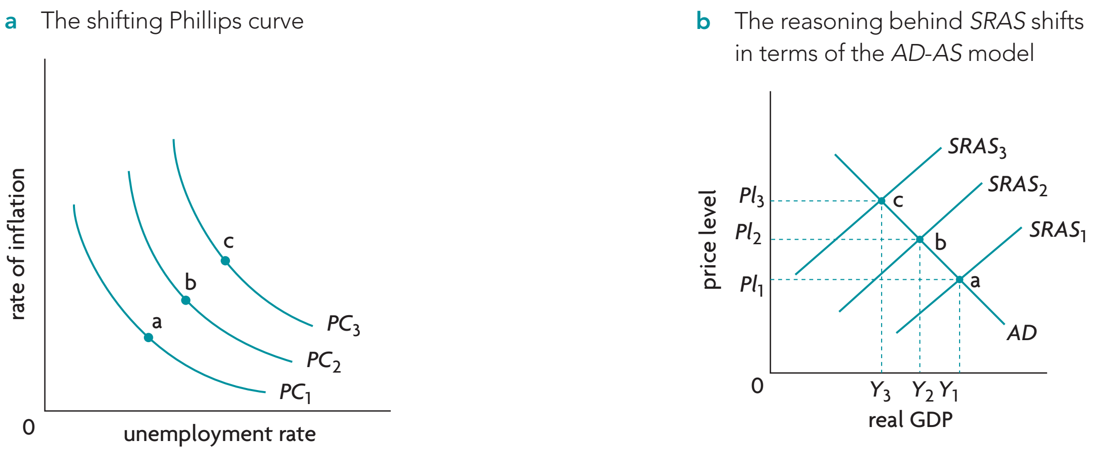


Figure 10.12: Stagflation: outward shifts of the short-run Phillips curve due to decreasing *SRAS*

The long-run Phillips curve and the natural rate of unemployment

In the late 1970s, the Nobel Prize-winning, monetarist economist Milton Friedman attacked the idea of a stable negative relationship between inflation and unemployment, and argued that there is only a temporary trade-off between inflation and unemployment, not a permanent one. Friedman made a distinction between a short-run Phillips curve and a long-run Phillips curve.

The short-run Phillips curve is what we have considered in Figure 10.11(a), which we can see once again in Figure 10.13(a), represented by $SRPC_1$ and $SRPC_2$. According to Milton Friedman, in the long run, this negative relationship no longer holds. Instead, the long-run Phillips curve is vertical at the level of 'full employment', or where unemployment equals the *natural rate of unemployment*. (In fact, the 'natural rate of unemployment' is a concept first introduced by Milton Friedman.) The long-run Phillips curve is *LRPC* in Figure 10.13(a).

Why is the long-run Phillips curve vertical at the economy's natural rate of unemployment? The answer is quite simple: it is so for the same reasons that the *LRAS* curve is vertical at the level of real GDP corresponding to the natural rate of unemployment (this was explained in Chapter 9, Section 9.3). Consider Figure 10.13, and suppose the economy is initially at point a in both parts. (Note that Figure 10.13(b) is the same as Figure 9.8(b) in Chapter 9.) In part (b), point a indicates that the economy is at a point of long-run equilibrium on AD_1, $SRAS_1$ and the *LRAS* curve, with real GDP equal to potential GDP shown by Y_p. At Y_p, unemployment is equal to the

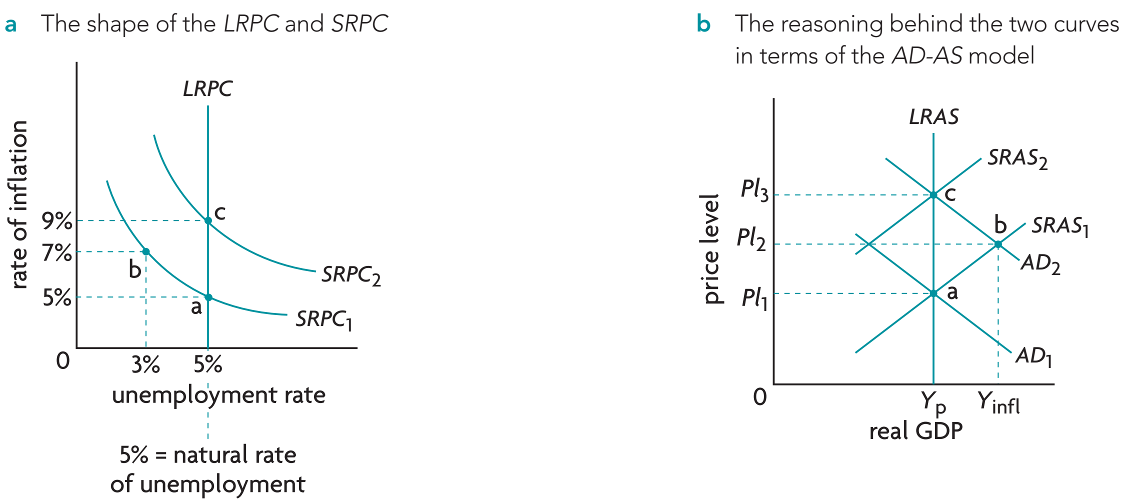


Figure 10.13: The short-run and long-run Phillips curves

natural rate of unemployment, which we assume to be 5%. In part (a), point a indicates that the economy is on a short-run Phillips curve, $SRPC_1$, where it is experiencing a rate of inflation of 5% and an unemployment rate of 5%, or the natural rate of unemployment.

Suppose there occurs an increase in aggregate demand, so that the *AD* curve in part (b) shifts from AD_1 to AD_2. In the short run the economy moves to point b on the $SRAS_1$ curve, corresponding to a higher price level, Pl_2, increased real GDP, Y_{infl}, and lower unemployment (unemployment falls below the natural rate). This corresponds to point b on the $SRPC_1$ in part (a), where there is a higher inflation rate of 7% and lower unemployment rate at 3%.

The economy moved to point b in the short run, because in the short run wages are constant; with the price level increasing, firm profitability increases, output increases and unemployment falls. But in the long run, point b cannot be a point of equilibrium, because, as we know from Chapter 9, wages will rise to meet the increases in the price level, causing the *SRAS* curve to shift leftward from $SRAS_1$ to $SRAS_2$, where it intersects AD_2 at a point on the *LRAS* curve, or point c. Point c in part (b) is associated with a higher price level Pl_3, but real GDP has fallen back to Y_p, and the rate of unemployment has returned to the natural rate. In part (a), these changes mean the economy has moved to point c, where the short-run Phillips curve has shifted to the right to $SRPC_2$ (remember, when the *SRAS* curve shifts leftward with a constant *AD* curve, the *SRPC* curve shifts rightward, as we saw in Figure 10.12). At point c, there is a higher rate of inflation, now standing at 9%, and unemployment has climbed back up to 5%, or the natural rate. The vertical line connecting a and c is the long-run Phillips curve (*LRPC*), situated at the natural rate of unemployment.[14]

Since the natural rate of unemployment occurs at long-run equilibrium, it is also known as *equilibrium unemployment*.

The short-run Phillips curve is a tool preferred by Keynesian economists, who see in this the possibility of using policies that focus on influencing aggregate demand to make choices about the rate of inflation and the rate of unemployment (and therefore the level of real GDP). By contrast, the long-run Phillips curve is an analytical tool preferred by monetarist/new classical economists, who are highly skeptical about the effectiveness of demand-side policies, and who use it to show that expansionary demand-side policies are more likely to result in inflation than to influence unemployment and real GDP. These economists prefer policies that focus on influencing aggregate supply. We will come back to these issues in Chapter 13.

According to the **short-run Phillips curve** in Figure 10.11(a), there is a negative relationship between the rate of inflation and the unemployment rate, suggesting that in the short run policy-makers can choose between the competing alternatives of low inflation or low unemployment by using policies that affect aggregate demand. The **long-run Phillips curve** is vertical at the natural rate of unemployment, indicating that unemployment is independent of the rate of inflation, and that policy-makers do not have a choice between the two competing alternatives. In the long run, the only impact of an increase in aggregate demand is to increase the rate of inflation, while the level of real output and unemployment remain unchanged at the natural rate of unemployment.

The short-run Phillips curve is questioned once again

In the years following the global financial crisis, the short-run relationship between inflation and unemployment shown in Figure 10.11(a) came to be questioned again. Many economists around the world began to argue that this relationship has broken down. The reason is that while unemployment has fallen to very low levels, inflation has not been

[14] This same argument is often made in terms of actual and expected rates of inflation. Let's assume that when the economy is initially at point a, nominal wages are set on the expectation that the rate of inflation will be 5%, and therefore nominal wages have been agreed with employers to increase by 5% so as to maintain a constant real wage. Let's say that the increase in aggregate demand, however, in actual fact gives rise to an inflation rate of 7%. Real wages decline as a result, firm profitability increases, real GDP increases as the economy moves upward along $SRAS_1$, and unemployment falls below the natural rate to 3%. Thus we have the movement from point a to point b on $SRAS_1$ and on $SRPC_1$. In the long run, the economy moves to point c because nominal wages adjust to actual rates of inflation, with the result that real wages increase to their previous level, firm profitability falls to its original level, real GDP falls to Y_p, and unemployment climbs back to the natural rate of 5%. The only difference from the initial equilibrium is that there is now a higher rate of inflation, of 9%. In the long run, when the actual rate of inflation is equal to the expected rate of inflation, nominal wages increase in line with the actual rate of inflation, real wages remain constant, and the trade-off between inflation and unemployment disappears.

increasing as the Phillips curve would predict. A number of arguments have been put forward trying to explain this. According to one, rising global competition makes it difficult for firms to raise prices even as unemployment falls. According to another, wages have not been rising in many economically more developed countries, therefore there has not been a strong upward pressure on prices. Reasons why wages have not been rising include diverse factors like the decline of labour unions, technology, and globalisation. Still other economists argue that over the past seven decades when the Phillips curve began to be used as a basis for policy, there have been several occasions when the relationship between inflation and unemployment became unstable, as though it were taking a break from the normal pattern; this could again be a temporary break before the normal pattern resumes again. A debate has emerged, with some economists arguing that there is no longer any Phillips curve, while others are suggesting that it is only a matter of time before the inflation and unemployment relationship traced out in Figure 10.11(a) emerges once again.

TEST YOUR UNDERSTANDING 10.8

1. Explain why there may be a conflict between the goals of low unemployment and low inflation. (HL only)
2. a Using the concept of the short-run Phillips curve and the *AD-AS* model, explain why many economists during the 1960s considered that policy-makers had a choice between low inflation and high unemployment, or high inflation and low unemployment.

CONTINUED

   b Identify the events of the 1970s and 1980s that made economists believe that the short-run relationship between inflation and unemployment was unstable (not fixed and permanent).

   c Explain, using diagrams and the concept of stagflation, the relationship between shifts in the *SRAS* curve and the position of the short-run Phillips curve.
3. a Using one or more diagrams, show how the long-run Phillips curve differs from the short-run Phillips curve.

   b Outline what the long-run Phillips curve tells us about the relationship between the rate of inflation and the rate of unemployment in the long run.
4. Outline the relationship between

   a the long-run Phillips curve and the natural rate of unemployment, and

   b the long-run Phillips curve and the full employment level of output.
5. According to the theory of the Phillips curve, explain what will happen to the rate of inflation, the rate of unemployment and real GDP if policy-makers attempt to increase aggregate demand in order to increase the level of real GDP

   a in the short run, and

   b in the long run.
6. Using an *AD-AS* diagram, explain why the long-run Phillips curve is vertical.

THEORY OF KNOWLEDGE 10.2

Choosing between low unemployment and low inflation: the role of politics and ideology in economic policy

We return to the question posed at the end of the Theory of knowledge 10.1 earlier in this chapter: what is so important about measuring the natural rate of unemployment? In addition, how may the choice of policy goals be affected by the general political mood and ideology of societies and their governments?

Based on the Phillips curve analysis, we can easily answer the first question. If the actual rate of unemployment is above the natural rate, policy-makers can use demand-side policies to increase aggregate demand, without fearing inflation. If, however, actual unemployment is at or below the natural rate, any increase in aggregate demand only temporarily lowers unemployment, as this will go back to the natural rate once wages have adjusted, only at a higher price level (see Figure 10.13). In the long run, the increase in aggregate demand only creates inflation. Therefore, knowing the natural rate is important as a guide to policy-makers.

However, if the natural rate changes often, and cannot even be accurately estimated, there may be serious doubts about how reliable it is as the basis for guiding policy. Yet, since the 1970s, Friedman's thinking has been highly influential in creating a policy approach in many countries that focuses on keeping inflation low, even if unemployment is high. The argument is that *since demand-side policies to change aggregate demand cannot lower unemployment anyway, policy should focus on keeping inflation low.*

Many economists disagree with this perspective. According to Nobel Prize-winning economist Joseph Stiglitz:

> *'Policies that focus exclusively on inflation are misguided . . . As a practical matter, . . . the relationship between unemployment and inflation is highly unstable. It is virtually impossible to discern the relationship from the data except in a few isolated periods.*
>
> *[Policy-makers] face considerable uncertainty about the level of the [natural rate of unemployment]. Thus, they still face a trade-off between pushing unemployment too low, and setting off an episode of inflation, and not pushing hard enough resulting in an unnecessary waste of resources.*
>
> *How one views these risks depends on the costs of undoing mistakes . . . The weight of the evidence indicates that the cost of undoing the mistake of pushing unemployment down too far is itself very low . . . In this view, [policy-makers] should aggressively pursue low unemployment, until it is shown that inflation is rising.*
>
> *By contrast, inflation 'hawks'[15] argue that inflation must be attacked [preventively] . . . [T]his stance is a matter of religion, not economic science. There is simply little or no empirical evidence that inflation, at the low to moderate rates that have prevailed in recent decades, has any significant harmful real effects on output, employment, growth or the distribution of income. Nor is there evidence that inflation, should it increase slightly, cannot be reversed at a relatively minor cost . . .*
>
> *The view [that nothing can be done about unemployment] belongs to a school of modern macroeconomics that assumes . . . perfectly competitive markets . . . Because markets [in this view] are always efficient, there is no need for government intervention. More [dangerously], many supporters of this view, when confronted with the reality of unemployment, argue that it arises only because of government-imposed rigidities and trade unions. In their 'ideal' world without either, there would, they claim, be no unemployment.'*[16]

[15] Inflation hawks are policy-makers who believe that inflation has highly negative effects and should be controlled.

[16] Joseph Stiglitz (2006) *The Phelps Factor*, Project Syndicate.

CONTINUED

The idea that control of inflation is more important than keeping unemployment low is politically conservative, and is often embraced by people who believe in the superiority of free markets over government intervention to solve economic problems. Less conservative economists and workers who have only their job to rely on as a source of income, tend to prefer low unemployment over low inflation (provided of course that inflation is moderate and does not get out of hand). They also tend to favour some intervention in markets aiming to keep unemployment low.

The natural rate concept, favouring low inflation over low unemployment, became attractive to policy-makers for two reasons. One was that because of stagflation, it forced economists to question the Keynesian use of demand-side policies to deal with economic fluctuations. The second, and very important reason, was that since the late 1970s, there occurred a shift in the general political mood *away from government intervention and toward the market* (particularly in the United States and United Kingdom); the natural rate concept with its strong free-market orientation, offered itself as an appealing theoretical approach to policy-making. Its free-market recommendations, including abolishing or reducing minimum wages and reducing labour union power, were attractive to policy-makers who opposed intervention in markets. Therefore, the natural rate concept was conveniently adopted as a guide to policy, placing a greater emphasis on controlling inflation rather than reducing unemployment.

Thinking points

- What does Stiglitz mean when he says the perspective of inflation hawks 'is a matter of religion, not economic science'?
- On the basis of what knowledge criteria have societies made a consistent choice over many years to make a priority of low inflation over low unemployment?
- Why do many economists consider the policy choice between low inflation and low unemployment to be a battle between conservative and non-conservative economists?

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1. Research and find data on unemployment over a period of about ten years in selected countries you are interested in, and try to find a country that has experienced significant fluctuations in its unemployment rate. Try to find information on the causes of the fluctuations, specifically what are the factors that have led to increases or decreases over the years? Is it possible to come to any conclusions about whether cyclical fluctuations (the business cycle) have played a role in unemployment in these fluctuations? Try to find an example of structural unemployment, using the causes of structural unemployment you have learned about in this chapter.
2. Research and find data on inflation over a period of about ten years in the country you live in or a country you are interested in. You are likely to find alternating rising or falling inflation rates. Try to discover the causes of rising inflation or disinflation. See if you can determine whether these are demand-pull or cost-push factors.

CONTINUED

3 (HL only) In the text above it was noted that the short-run Phillips curve has come to be questioned by many economists, though many others argue that it is only a matter of time before the familiar inverse relationship between inflation and unemployment resumes. Research the most recent writings on the Phillips curve, in order to discover if and how the debate has been resolved. Determine whether unemployment and inflation are displaying the relationship shown by the short-run Phillips curve, and if they are not examine the reasons put forward to explain this.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 11

Macroeconomic objectives II:

Economic growth, sustainable level of debt

BEFORE YOU START

1. We have seen that countries tend to achieve economic growth over a period of time. What do you think might be some positive and negative consequences of economic growth?
2. Governments sometimes accumulate debt in order to achieve their economic objectives. What might be some problems faced by countries with high levels of debt?

In this chapter we will study two more important macroeconomic objectives: economic growth and sustainable level of debt.

Symbolising sustainable growth

11.1 Economic growth

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain short-term growth in terms of (AO2)
  - actual growth in the PPC model
  - the role of aggregate demand in the *AD-AS* model
- explain long-term growth in term of (AO2)
  - shifts in the PPC illustrating growth in production possibilities
  - the role of *LRAS* in the *AD-AS* model
- draw diagrams of the PPC model showing actual growth and growth in production possibilities (AO4)
- draw diagrams of the *AD-AS* model illustrating (AO4)
  - increases in *AD* to show increases in real output
  - increases in *LRAS* to show increases in full employment output
- explain how economic growth is measured (AO2)
- calculate the rate of growth from data (AO4)
- discuss the consequences of growth including its impact on (AO3)
  - living standards
  - the environment
  - income distribution

The meaning and significance of economic growth

You were introduced to *economic growth* in Chapter 1 in connection with the production possibilities model. It was also briefly discussed in Chapter 9 in connection with the *AD-AS* model.

Economic growth refers to an increase in real GDP, or the real quantity of goods and services produced over a period of time (typically a year), and is usually expressed as:

- a percentage change in real GDP over a period of time, or
- a percentage change in real GDP *per capita* over a period of time.

The distinction between growth in real GDP and real GDP *per capita* is important because the two measures are very likely to be different. Whereas real GDP measures the total output produced in an economy, real GDP *per capita* measures total output per person. Real GDP *per capita* is a better indicator of the standard of living of a population, since it measures the amount of real GDP corresponding to each person on average.

Note that economic growth is measured as the percentage change in *real* as opposed to *nominal* GDP. The real figures eliminate the impact of price changes, thus permitting meaningful comparisons of levels of output over time.

Calculating economic growth

Calculating economic growth from a set of data

The formula for calculating percentage change in real GDP is the following:

% change in real GDP =

$$\frac{\text{final value of real GDP} - \text{initial value of real GDP}}{\text{initial value of real GDP}} \times 100$$

For more information on percentage changes, see 'Quantitative techniques' chapter in the digital coursebook.

If an economy had real GDP of \$75.3 billion in 2017 and \$81.7 in 2018, by how much did real GDP grow in 2017–2018? Using the formula above we find

$$\%\text{ change in real GDP} = \frac{81.7 - 75.3}{75.3} \times 100$$

$$= \frac{6.4}{75.3} \times 100 = 8.5\%$$

The identical formula is of course used to calculate the percentage change in real GDP *per capita*. Suppose an economy has real GDP *per capita* of \$1402 in 2017,

$1457 in 2018 and $1410 in 2019. By how much did real GDP *per capita* grow in 2017–2018 and in 2018–2019?

2017–2018 change in real GDP *per capita*

$$= \frac{1457 - 1402}{1402} \times 100$$

$$= \frac{55}{1402} \times 100 = 3.9\%$$

2018–2019 change in real GDP *per capita*

$$= \frac{1410 - 1457}{1457} \times 100$$

$$= \frac{-47}{1457} \times 100 = -3.2\%$$

Note that in the second period, economic growth was negative.

Relating growth in real GDP to growth in real GDP per capita

Suppose real GDP is growing in a hypothetical economy, so it has a positive growth rate. Does this mean it also has positive *per capita* GDP growth? The answer depends on how fast the population is growing. If real GDP is growing faster than the population, then the amount of real GDP that corresponds to each person on average increases, resulting in positive growth in real GDP *per capita*. If, on the other hand, the population is growing faster than real GDP, then the amount of GDP per person on average decreases, and the growth rate of real GDP *per capita* is negative.

If we know the percentage change in real GDP and the percentage change in the population, we can find the percentage change in real GDP *per capita* in a very simple way:

% change in real GDP *per capita* = % change in real GDP − % change in population

For example, if real GDP grew by 2% in a year, and the population grew by 1.5%, then real GDP *per capita* growth was 0.5%. If, however, the population grew by 2.5%, then the % change in real GDP *per capita* was −0.5%, indicating that output per person fell.

The significance of economic growth

Economic growth rates achieved by countries around the world vary widely. While some countries experience rapid growth, others grow much more slowly, while others contract for a period of time. This can be seen in Table 11.1, showing average annual growth rates of real GDP for several countries for the period 2013–2018.

	Average annual real GDP growth (%) 2013–2018
Ethiopia	9.40
Bangladesh	6.72
Indonesia	5.51
Sweden	2.71
Japan	1.25
Argentina	0.15
East Timor	–4.81

Source: *https://en.wikipedia.org/wiki/List_of_countries_by_real_GDP_growth_rate*

Table 11.1: Real GDP percentage growth

Differing growth rates have enormous implications for a country's economic performance over long periods of time. Let's consider what would happen to real GDPs of three hypothetical economies that grow at different rates over a period of five years (as in the table). Imagine that each economy starts out in the year 2015 with a real GDP of $1000. The first one grows for five years at the high annual rate of 9.40%; the second country grows at the low average annual rate of 0.15 %; and the third country contracts at the annual rate of –4.81% (negative growth). The results are presented in Table 11.2.

The country that grows at 9.40% per year for five years ends up with a real GDP that is double that of the country with the negative growth rate of –4.81%. The country that grows at the low average rate of 0.15% per year only adds $8 to its GDP.

2015 GDP	Annual growth rate	2020 GDP
$1000	9.40%	$1567
$1000	0.15%	$1008
$1000	–4.81%	$782

Table 11.2: Growth of real GDP in hypothetical economies

The large cumulative impact that rates of growth have on levels of real GDP explains why governments around the world focus strongly on policies intended to increase their growth.

TEST YOUR UNDERSTANDING 11.1

1 Describe the meaning of economic growth and explain its significance over a period of years.

2 When we calculate economic growth, outline whether we should use nominal or real values of GDP.

3 Explain the advantage of calculating growth of real GDP in *per capita* terms rather than growth of real GDP.

4 Suppose an economy had real GDP *per capita* of €1579 in 2017, €1611 in 2018 and €1597 in 2019. Find the rate of economic growth

a in 2017–2018, and

b in 2018–2019.

c State when the economy experienced negative growth.

5 Outline how it is possible that a country can have positive real GDP growth and yet have negative real GDP *per capita* growth.

6 Suppose that an economy's real GDP grew by 2.2% in 2007, and its population grew by 1.5% during the same year. Calculate by how much its real GDP *per capita* grew.

Short-term growth versus long-term growth

We will now make a distinction between **short-term growth**, which takes place over relatively short periods of time, and **long-term growth**, which needs a long time to take effect.

Many of the ideas that will be discussed in this section have been studied in previous chapters, and so much of what you will read about in this section will be a review for you.

Understanding growth using the *AD-AS* model

In the *AD-AS* model, economic growth, or increases in real GDP, occurs as a result of:

- increases in aggregate demand, which is referred to as short-term growth
- increases in short-run aggregate supply (though this is far less common), also a type of short-term growth
- increases in long-run aggregate supply, referred to as long-term growth.

Short-term growth: increases in aggregate demand (*AD*)

Figure 11.1 shows short-term growth in the monetarist/new classical model. It can be caused by increases in aggregate demand, illustrated in part (a) by the rightward shift of the *AD* curve from AD_1 to AD_2, resulting in a real GDP increase from Y_1 to Y_2. Note that *short-term economic growth does not involve an increase in potential output*, and therefore there is no rightward shift of the *LRAS* curve.

In the Keynesian model, short-term economic growth can be seen in Figure 11.1(b), where successive increases in aggregate demand from AD_1 to AD_2 to AD_3 and AD_4 result in growth of real GDP from Y_1 to Y_2 to Y_3 and Y_4. Note that here, too, *short-term economic growth does not involve an increase in potential output*, and hence no rightward shift of the *AS* curve.

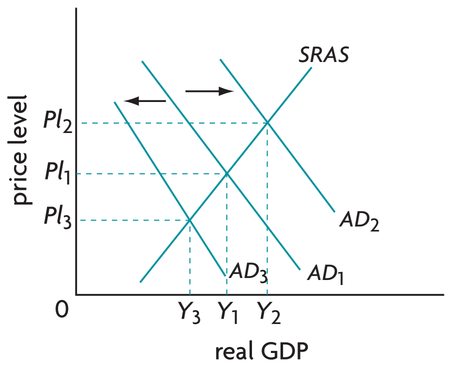
This type of short-term growth can be caused by any of the factors that can cause increases in aggregate demand (*AD*) that you studied in Chapter 9, Section *Determinants of aggregate demand. AD* can increase as a result of an increase in any one or more of the components of *AD*, which are consumption (*C*), investment (*I*), government spending (*G*) and net exports ($X - M$).

Short-term growth: increases in short-run aggregate supply (SRAS)

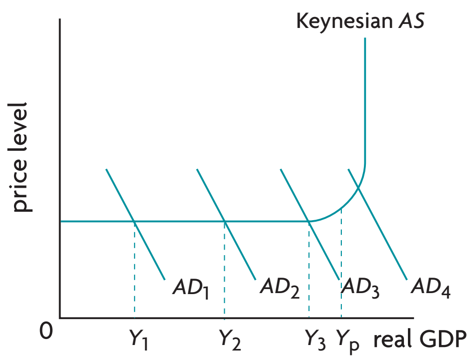
Short-term growth can also be caused by increases in short-run aggregate supply, or a rightward shift of the *SRAS* curve from $SRAS_1$ to $SRAS_2$ in Figure 11.1(c), causing real GDP to increase from Y_1 to Y_2. The causes of this type of growth include the factors that can increase short run aggregate supply (*SRAS*) that you studied in Chapter 9, Section *Changes in short run aggregate supply*: falls in prices of factors of production (labour and non-labour resources), increases in subsidies, or positive supply shocks.

It should be noted, however, that short-term economic growth is affected far more by increases in aggregate demand rather than in short-run aggregate supply.

a The monetarist/new classical model: increase in aggregate demand



b The Keynesian model: increase in aggregate demand



c The monetarist/new classical model: increase in short-run aggregate supply

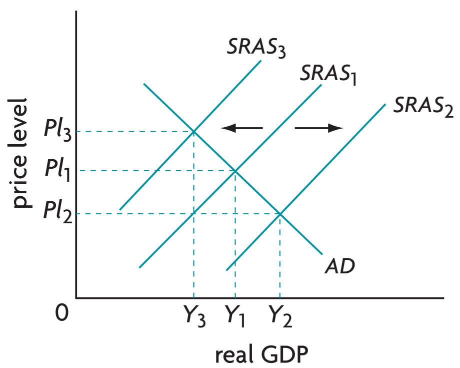


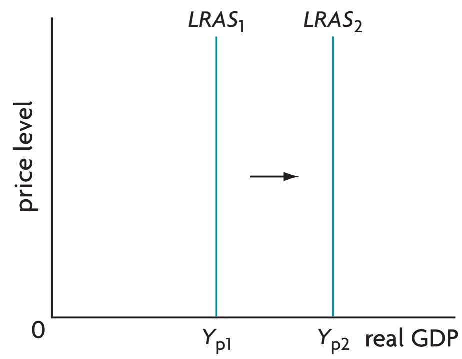
Figure 11.1: Short-term growth

Long-term growth: increases in long-run aggregate supply (*LRAS*) or Keynesian *AS*

In Chapter 9, Section *Shifting aggregate supply curves over the long term*, you learned about a number of factors that affect the positions of the *LRAS* and Keynesian *AS* curves, causing these curves to shift to the right. Since these usually need an extended period of time to take effect, their effects on the economy are referred to as *long-term growth*. Both of these are shown in Figure 11.2, which is the same as Figure 9.13.

As you may recall, these factors include increases in the quantity and improvements in the quality of factors of production, technological change, improvements in efficiency and institutional changes.

a The monetarist/new classical model



b The Keynesian model

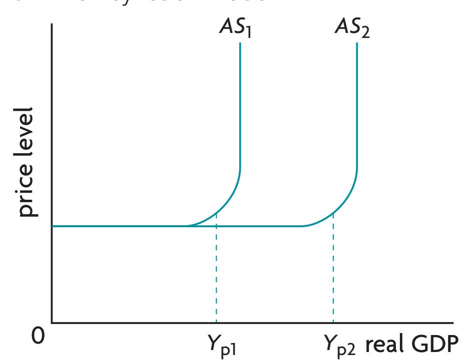


Figure 11.2: Increasing potential output, shifts in aggregate supply curves and long-term economic growth

The *AD-AS* model, growth and the business cycle

In Chapter 9 we saw that inflationary and deflationary gaps correspond to the upward and downward phases of the business cycle introduced in Chapter 8, Figures 8.4 and 8.5. We also saw that when there is no inflationary or deflationary gap, the economy produces at its full employment level of output, or potential GDP. It is now simple to see the connection between economic growth and the business cycle.

Short-term growth is shown in the expansion phase of the business cycle, when real GDP is increasing over time, though this is followed by a contraction, or a period of decreasing real GDP. The contraction illustrates situations where there is negative growth. The upward phases of the business cycle are caused mainly by increases in aggregate demand, and to a lesser extent by increases in short-run aggregate supply. Similarly the downward phases, illustrating negative growth, are caused by decreasing aggregate demand and to a lesser extent decreasing short-run aggregate supply.

On the other hand, over long periods of time, most economies experience positive economic growth. The business cycle diagram, showing an upward-sloping, long-term growth trend, indicates that the economy's real GDP is growing over time. The long-term growth trend, you may remember, was referred to as *potential output*; this is none other than the real GDP level at which the *LRAS* curve is situated, and the potential output we see in the Keynesian model.

> Short-term growth caused mainly by increases in aggregate demand (and to a lesser extent increases in short-run aggregate supply) corresponds to expansions of real GDP in the business cycle diagram. Long-term growth caused by rightward shifting *LRAS* or Keynesian *AS* curves that show increases in potential output corresponds to the long-term growth trend in the business cycle diagram.

Figure 11.3 illustrates how macroeconomic equilibrium changes over the long term when potential output is increasing. Y_1 and Y_2, which are the long-run equilibrium points of part (a) and the Keynesian equilibrium points of part (b) correspond to points in the business cycle diagram where actual output is equal to potential output. These are the points of intersection of the curves showing actual GDP and potential GDP. The long-term growth trend of the business cycle diagram simply traces out the increases in potential output shown in the *AD-AS* models.

In the real world, it is very difficult to examine economic activity and arrive at accurate conclusions about what part of growth is due to short-term fluctuations of the business cycle and what part to growth in potential output. However, economists continuously make efforts to measure potential output and its growth, because having estimates of these can help governments formulate appropriate policies to guide the macroeconomy in the desired directions.

Understanding growth using the production possibilities model

This model was introduced in Chapter 1 (Section *Introducing the production possibilities model*). It is suggested that you reread this section as it is highly relevant to the discussion of growth below.

Short-term growth: growth in actual output

You may remember that the production possibilities curve (*PPC*) shows combinations of maximum output that can be produced by an economy with fixed resources and technology, *provided there is full or maximum employment of resources and efficiency in production*. Maximum employment in this model does not mean 'full employment' as in *AD-AS* models; it means that all resources are employed to the fullest extent and there is zero unemployment.

In Chapter 1, we learned that it is highly unlikely for any economy to be producing on its *PPC*, as this presupposes full or maximum employment of all resources and efficiency, which cannot be achieved in the real world. Any country is therefore most likely to be producing at a point inside its *PPC*. It can move closer to its *PPC* and *increase the actual quantity of output it produces by*

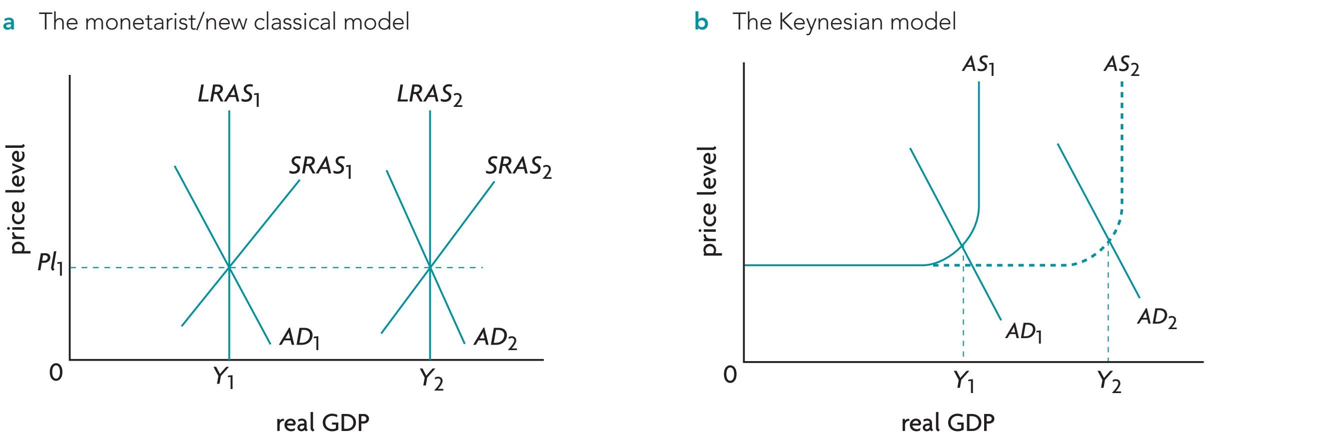


Figure 11.3: Long-term economic growth: achieving potential (full employment) output in a growing economy

reducing unemployment or by improving the efficiency of resource use. In Figure 11.4(a), which is the same as Figure 1.3(a) in Chapter 1, the movement from point A to point B shows growth of actual output. We refer to this as *actual growth*, which is a kind of *short-term growth*, because it can occur over short periods of time.

Long-term growth: growth in production possibilities

It is clear from Figure 11.4(a) that reduction of unemployment and inefficiencies can only result in a limited amount of economic growth. The presence of the *PPC* sets an upper bound to how much more output can be produced. The only way to produce more output beyond the limit created by the *PPC* is if the *PPC* shifts outward as in Figure 11.4(b), allowing more of both groups of goods (*military goods* and *consumer goods*), to be produced. The shifts from PPC_1 to PPC_2 to PPC_3, called *growth in production possibilities*, are referred to as *long-term growth*, because it is likely to occur over long periods of time. In Figure 11.4(b), together with the growth in production possibilities there is also actual growth shown by the outward movements of the economy's actual output, from A to B to C.

As you learned in Chapter 1, the factors that lead to outward shifts of the *PPC*, or increases in production possibilities are:

- increases in the quantity of resources (factors of production) in the economy
- improvements in the quality of resources (for example, through more educated labour, or improved physical capital through technological change).

As production possibilities grow, efforts must be made to keep unemployment at low levels and reduce inefficiencies to ensure that actual output grows along with production possibilities, as in Figure 11.4(b). For example, if the size of the labour force increases, this will not lead to actual growth if much of this labour remains unemployed, in this case the economy could remain stuck at point A even as PPC_1 shifts to PPC_2. Similarly, the discovery of major oil reserves may not lead to actual growth if these reserves remain unexploited, or if their exploitation is undertaken inefficiently.

As you may remember from Chapter 1, the *PPC* can also shift inward, indicating a decrease in production possibilities. This means that less of the two goods is being produced, as shown in Figure 11.4(c). This results from a decrease in the quantity of resources or worsening of resource quality. It represents the case of negative growth.

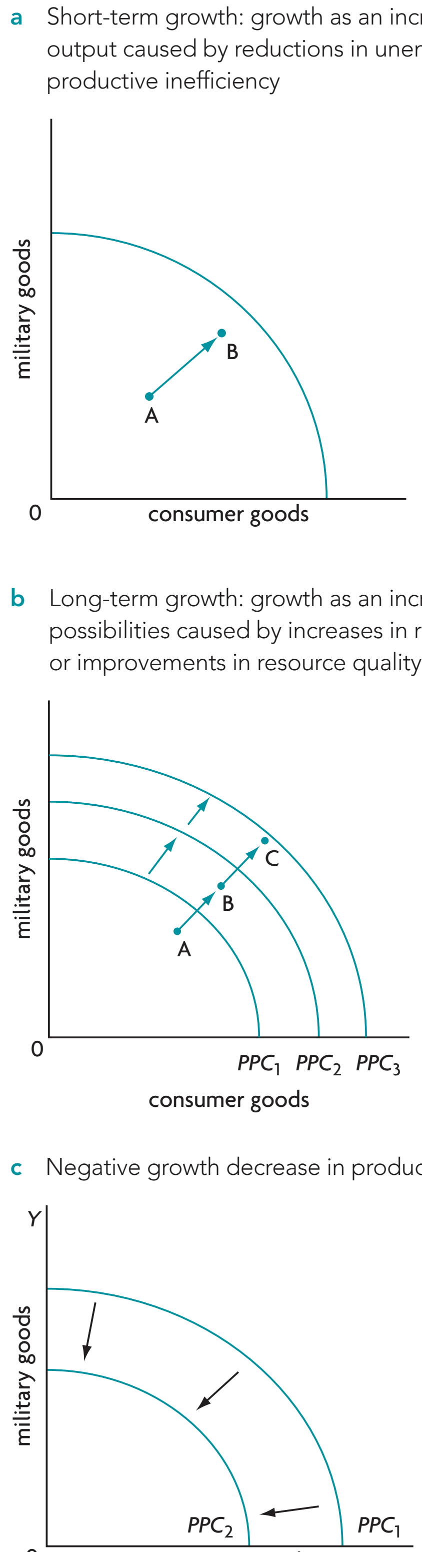


Figure 11.4: Using the production possibilities model to illustrate economic growth

> Actual growth is *short term-growth*, because it can occur over short periods of time and is due to reductions in unemployment or inefficiency in production. Growth in production possibilities is *long-term growth*, because it usually requires long periods of time and is due to increases in quantity or improvements in quality of factors of production.

Table 11.3 provides a summary of the factors that cause short-term and long-term growth in the two models. We can see that they are in some respects similar regarding the causes of growth. In the *PPC* model, reduction of unemployment is a key cause of actual growth over the short term. In the *AD-AS* model, increases in *AD* or *SRAS* both involve reductions in unemployment, since it is not possible to produce a larger amount of real GDP without more labour being hired, which leads to a fall in unemployment. Therefore, in both models falls in unemployment are a major factor behind growth.

Regarding long-term growth, we can see that the first three causes of growth that are listed are common to both models. The similarities between the two models are hardly surprising since both illustrate economic growth. On the other hand it is also not surprising that there are some differences between the two models. The two *AS* curves are intended to show the relationship between the price level and real GDP. The production possibilities model does not deal with the price level, it is better suited to illustrate the principles of resource scarcity and opportunity costs of economic choices.

<table>
<tr><th></th><th>AD-AS model</th><th>PPC model</th></tr>
<tr><td>Short-term growth</td><td>• increases in AD
• increases in SRAS (less important)</td><td>• reduction in unemployment
• improvement in efficiency</td></tr>
<tr><td rowspan="2">Long-term growth</td><td colspan="2">• increased resource quantity
• improved resource quality
• technological change</td></tr>
<tr><td>• improvements in efficiency
• institutional changes</td><td>–</td></tr>
</table>

Table 11.3: Factors that cause economic growth

TEST YOUR UNDERSTANDING 11.2

1. Using diagram(s), explain the difference between an increase in actual output and an increase in production possibilities. Using examples identify the causes of each.
2. a Using diagrams, explain the difference between long-term economic growth and short-term economic growth due to economic fluctuations.

   b Using examples, identify the causes of each.

   c Outline how potential output changes/does not change in each case.
3. Use the production possibilities model and diagrams to show how the following can result in economic growth (positive or negative), distinguishing between *PPC* shifts and movements of a point closer to or further from a given *PPC*:

   a a discovery of new oil reserves,

   b a fall in natural unemployment,

   c an increase in cyclical unemployment,

CONTINUED

d an improvement in levels of health of the population,

e an improvement in efficiency in production,

f the widespread use of a new technology,

g a violent conflict destroys a portion of a country's factories, machines and road system,

h large cuts in government spending on education and health care lower levels of education and health in a population,

i an increase in the quantity of capital goods,

j an improvement in the level of education and skills of workers, and

k industrial pollution destroys the environment.

Why economic growth occurs (Supplementary material)

The roles of physical, human and natural capital in economic growth

To understand why increases in resource quantities and improvements in resource quality including technological change cause long-term growth in both models, we will use the expanded meaning of *capital*, introduced in Chapter 1, Section *Resources as factors of production*. 'Capital' generally refers to resources that can produce a future stream of benefits. This future stream of benefits arises from investment, or spending undertaken to create that stream of benefits.

Three of the four factors of production: land, or natural resources; labour, or human resources; and physical capital, can be interpreted as a type of 'capital':

- *Physical capital,* also referred to as 'capital goods' is the standard type of capital; as we know from Chapter 8 it results from investments, or spending to produce machines, equipment, roads, etc.
- *Human capital* refers to the skills, abilities, knowledge and levels of health of workers. Human capital results from investments, or spending on education, training, provision of health care services, clean water supplies, good nutrition, and generally anything that affects levels of education and health.
- *Natural capital* includes everything that traditionally falls within 'land', or 'natural resources'. It includes everything under the land (mineral deposits, metals, oil, natural gas, etc.) plus everything on the land (rivers, lakes, oceans, forests, soils, etc.), plus a country's overall natural environment and ecosystem (air, wildlife, biodiversity, climate, ozone layer, and so on). Whereas 'land' and 'natural capital' include the same things, there is an important difference in how the two terms are interpreted. 'Land' is assumed to be given by nature and does not change. 'Natural capital' does change, because it can be destroyed (by cutting down forests, polluting the air and water, depleting fish). It can also be improved (by planting more forests, improving soil quality). Maintaining natural capital and improving its quality depend on investments that aim to preserve and improve natural resource quantity and quality.

By expanding the meaning of 'capital', we have extended the meaning of 'investment' to apply to all three factors of production listed above:

Factor of production	Type of capital
physical capital	is related to investments in physical capital
labour = human resources	is related to investments in human capital
land = natural resources	is related to investments in natural capital

Investments can be undertaken by the private sector (firms or private individuals) or by the public sector (the government).[1] *Investment is crucial to building capital of any type.*

[1] Note that in the *AD-AS* model, private investment is carried out by firms (*I*) and government investment is included in the *G* component.

We will now see how investments in the three types of capital lead to economic growth.

Physical capital, technology and economic growth

An increase in the *quantity of physical capital* involves an increase in the number of machines, tools, equipment, road systems, ports, etc. available in an economy.

An improvement in the *quality of physical capital* depends on technological advances, which lead to new and better machines, tools and equipment. Technological advances are usually incorporated into new capital goods; for example, a new computer that is faster and more powerful incorporates within it the new technology that makes it faster and more powerful. When a technological advance is incorporated into a capital good, it is referred to as being *embodied* in the new capital.

Therefore, 'improved capital goods' are capital goods that embody a new technology. Use of capital goods embodying new technologies leads to a larger quantity of output produced; for example, the use of a more powerful computer allows a worker to produce more output.

> Increases in the quantity and improvements in the quality of physical capital, arising from investments in physical capital and new technology, are among the most important sources of economic growth over long periods of time.

Human resources, human capital and economic growth

The quantity of labour can sometimes be an important source of economic growth; for example, the influx of foreign workers into Germany in the 1960s and 1970s played an important role in promoting its growth. However, many countries (especially less developed ones) sometimes face high levels of unemployment and underemployment. Therefore, increases in the quantity of labour may not always be a source of growth.

Far more important than increases in the quantity of labour is improvement in the quality of labour, determined by skills, abilities, knowledge and levels of health of the workforce. Improved labour quality is the result of investments in human capital, including spending on education, building schools, providing meals for schoolchildren and providing vocational training; as well as spending to provide medical services, immunisation, and ensuring access by the overall population to health care services, together with the provision of sanitation and clean water supplies, and keeping the environment unpolluted.

Higher levels of skills, knowledge and health, resulting from investments in human capital, are a very important source of economic growth because a highly skilled, well-educated and healthy labour force is more productive than an unskilled, uneducated and unhealthy one: a skilled and healthy worker can produce more output than a worker who is unskilled or unhealthy.

> Increased quantities of labour are unlikely to be a source of economic growth over long periods, but improvements in the quality of labour, arising from investments in human capital, are among the most important sources of growth.

Natural resources, natural capital and economic growth

When thinking about the contribution of natural resources (natural capital) to economic growth, it is useful to make a distinction between two kinds of natural capital: *marketable commodities* (commodities that are bought and sold) such as timber, minerals, metals, natural gas, coal and oil; and ecological resources such as soil quality, rivers, clean air, biodiversity, the ozone layer (and, more generally, common pool resources).

The role of marketable commodities

Marketable commodities can contribute to growth but are not essential. For example, the United States benefited enormously from its large tracts of good quality agricultural land, oil reserves and mineral deposits. Yet the evidence suggests that countries do not need to be rich in marketable commodity-type natural resources to achieve high rates of growth. There are many economies, such as Israel, Japan, Hong Kong, Singapore, South Korea, Switzerland, Taiwan and others, that have achieved high rates of growth over long periods and have attained high levels of GDP *per capita*, in spite of producing few, if any, marketable commodities.

The role of common pool resources

Common pool resources are crucially important to long-term growth. Long-term economic growth depends critically on the ability of countries to maintain, and if possible improve, environmental quality, and therefore natural capital that includes common pool resources (see Chapter 5). Environmental destruction can have direct effects on the amount of output produced; for example, a farmer working with poorer quality soils produces less output; fisheries in fish-depleted seas have a smaller catch. Despite this, it can also have important indirect effects; for example, workers whose health is affected by environmental pollution become less productive, and this involves depletion of human capital together with depletion of natural capital.

Analysed in terms of the production possibilities model, these effects involve an inward shift of the *PPC* due to fewer and lower quality environmental and human resources (natural and human capital) and therefore lower (or even negative) economic growth in the future, seen in Figure 11.4(c). Therefore, continued economic growth in the future requires investments in the present in natural capital for environmental preservation.

The central importance of productivity as a source of economic growth

The contributions of resource quantities and quality to economic growth can be summarised in the concept of *productivity*, referring to the quantity of output produced for each hour of work of the working population. For an economy as a whole, productivity can be measured as real GDP divided by the total number of hours worked.

An improvement in productivity means that workers become more productive: the quantity of output produced in an hour of work increases. Improvements in productivity lead to economic growth, because each hour of work now produces more output.

What are the factors that cause improvements in productivity? They are exactly the factors making the most important contributions to long-term economic growth discussed above. They include:

- increases in quantity and improvements in quality of physical capital (through investments in physical capital and technological change)
- improvements in the quality of labour (through investments in human capital)
- improvements in (or at least maintenance of) the quantity and quality of ecological resources (through investments in natural capital).

Productivity improvements result in rightward shifts of the *LRAS* or Keynesian *AS* curves, and outward shifts of the *PPC*.

TEST YOUR UNDERSTANDING 11.3

1 Explain the relationship between investment and physical, human and natural capital.

2 Referring to land, labour and physical capital, and the concept of capital as applied to each of these resources, explain how each one can contribute to economic growth.

3 a Define productivity.

  b Explain why improved productivity is important for economic growth.

  c What are the most important factors that result in productivity improvements?

  d Show productivity growth using *PPC* and *LRAS* diagrams.

Consequences of economic growth

Economic growth impacts upon many aspects of the economy, and some of its possible consequences are positive while others may be negative. It is important to note that many of these consequences, whether positive or negative, are not inevitable, *but rather follow from the ways that growth is pursued*.

Impact of economic growth on living standards

Living standards (or *standards of living*) refer to levels of income, wealth and consumption of goods and services, including health care and education. If real GDP of a country increases faster than its population, then an increase in GDP *per capita* results. This indicates that *there is a greater potential for people to increase their consumption of goods and services, and improve their standards of living*.

According to numerous studies carried out over many years, economic growth is associated with improvements in standard of living indicators. This is what we would expect, since growth provides additional resources allowing for improvements in living standards. However, GDP *per capita* or income *per capita* is only an average measure, and does not tell us how the increase in income is distributed or whether there is a broadly distributed improvement in living standards. Therefore improvements vary a lot from country to country and from time period to time period, so that for a given rate of growth they are in some cases small and in others much larger. What accounts for such differences?

Important factors that allow economic growth to have positive effects on standards of living include the following:

- **The distribution of income.** The greater the share of income going to poorer households, the greater the potential for contributing to improvements in living standards as the poorer households are those with the greatest deprivations. If increases in income made possible by economic growth bypass the poorer households, growth has limited effects on broadly shared improvements in living standards.
- **Household spending.** The greater the share of household income spent on goods and services such as food, education and health care, the greater the improvements in living standards.
- **The share of income controlled by women.** The greater this is, the stronger the impact (see also Chapters 19 on the role of women).
- **Government spending on merit goods.** This relates to the share of the government budget allocated to priority areas like education, health care and infrastructure including clean water supplies and sanitation; the larger this is, the greater the positive effects of growth.
- **Contributions by non-governmental organisations (NGOs).** Because of their poverty orientation and their general effectiveness in reaching poor people, NGOs contribute to increasing the impact of growth on higher standards of living.

A major study of data between 1970 and 2005 for 111 countries by the United Nations Development Programme (UNDP)[2] shows that the greatest improvements in literacy and life expectancy (two components of the Human Development Index (HDI); see Chapter 18) *are not occurring in the fastest growing economies of the world* (the only two exceptions being China and Korea). Factors contributing to HDI improvements are government expansion of education and health care, together with the international community's contribution of vaccines and antibiotics.

Therefore, *while economic growth offers the potential to achieve improvements in standards of living, these improvements do not occur automatically as a result of economic growth* but require appropriate policies to make effective use of the resources growth makes available.

Impact of economic growth on the environment

Experience shows that growth, especially rapid growth, often leads to unsustainable resource use (particularly in the case of common pool resources). For example, very high growth rates in East Asian countries have been associated with serious environmental losses taking the form of very high levels of urban air pollution, soil degradation due to soil erosion, waterlogging and overgrazing, threats to biodiversity and serious deforestation. Industrialisation based on fossil fuels is a major source of pollution (negative production externalities). Increasing incomes lead to consumption patterns also based on greater fossil fuel consumption (use of cars, air conditioners, air travel, etc., creating negative consumption externalities). Other activities, such as commercial logging and agricultural practices based on a lack of pricing mechanism for common pool resources, result in their unsustainable use.

Experiences like these have led to the widespread belief that economic growth and environmental sustainability are conflicting objectives: more of one means less of the other. Many governments around the world have based their policies on this belief by following the 'grow now, clean up later' way of thinking, which argues that since using resources to preserve the environment reduces growth, it is preferable to pursue growth with its negative effects on the environment, and postpone the 'clean-up' job of environmental preservation for later when incomes will be higher. For example, the installation of pollution-control equipment involves greater costs for firms, which may mean lower profits, lower investment and lower economic growth. Switching to environmentally sound agricultural practices similarly involves costs that may cut into future economic growth prospects. Setting limits to deforestation for timber places restrictions on the growth of the timber industry. Therefore, allocating resources for environmental protection arguably translates into smaller increases in output and hence lower economic growth.

[2] United Nations Development Programme, Human Development Report 2010.

Yet, this way of thinking is unsound for several reasons. One is that *some environmental damage is irreversible*; it will not be possible to correct the damage in the future, and some resources will be lost forever. For example, lost biodiversity can never be recovered; lost lives due to pollution-induced illnesses can similarly never be recovered. A second is that *it justifies government inaction on the environment*. Governments and policy-makers often wrongly assume that environmental issues will automatically be regained in the future as incomes increase with growth. This is unrealistic, because preservation of the environment requires policies aiming to limit negative environmental externalities.

A third, related reason is that *it is not growth itself that is bad for the environment, but rather the ways that growth is pursued*. If growth were pursued differently, it need not conflict with environmental sustainability. A fourth reason is that *growth based on unsustainable resource use may lead to destruction of natural resources on such a wide scale that the possibility of continued future growth may be threatened*.

Modern growth theory shows that economic growth and environmental sustainability are in fact consistent with each other, and *can be successfully pursued together under certain conditions*, such as the following:

- Governments implement market-based policies that 'internalise the externalities', thus not only correcting them (at least in part) but also providing incentives for sustainable resource use and promotion of green (or 'clean') technologies (see Chapter 5).
- Governments pursue more environmental regulations that encourage pollution-free technological change (green technologies).
- There is an increased emphasis on human capital in production (which is pollution-free) as opposed to physical capital.
- There is an increased emphasis on 'green' investments, which promote growth while not hurting the environment: building public transportation systems; investing in insulation in homes and buildings; investing in clean technology research and development (R&D) and clean technologies.
- There are changes in the structure of the economy toward more services (which tend to be pollution-free), together with more investments in the protection of natural resources.

As incomes increase with economic growth, more resources are made available with which governments can pursue the above kinds of policies, encouraging economic growth at the same time that they encourage sustainability. Therefore, *economic growth and sustainability can be pursued together provided governments take appropriate measures to ensure sustainable resource use*. This is the very meaning behind the concept of 'sustainable development' (see Chapter 5).

However, even under the best possible circumstances where all of the above conditions are fulfilled, modern growth theories show that *there is a maximum rate of growth that is consistent with environmental sustainability*, and that if an economy exceeds this rate, resource use will become unsustainable. The reason is that pursuit of sustainability uses up some resources (for example anti-pollution controls, costs of regulation, etc.), and these resources represent an opportunity cost in terms of lost economic growth. Note, however, that this only applies to a loss of a portion of very high rates of growth.[3]

THEORY OF KNOWLEDGE 11.1

The conflict between economic growth and sustainability

The apparent conflict between growth and sustainability arises because of the way *economic growth is defined*. The idea of conflict follows from the conventional measure of economic growth, taken to be increases in real GDP *per capita*. This in turn depends on the definition of GDP, which is the value of all goods and services produced in a country in the course of a year.

As you may remember from Chapter 8, Section *Evaluating national income statistics*, one of the serious limitations of the measure of real GDP per capita is that it does not account for negative externalities

[3] You may be wondering what these rates of growth are. It is not possible to attach numerical values to these given the present state of economists' and scientists' knowledge, because not enough is known about the costs and benefits of growth and environmental sustainability.

CONTINUED

and environmental losses arising from environmental degradation. This limitation could be corrected if countries adopted other accounting methods that included environmental degradation in some form (such as the OECD Better Life Index or the Happy Planet Index). If economic growth was calculated in a way that took account of environmental destruction as well as possible negative health consequences of pollution we would find that in most countries around the world economic growth rates would be far lower than those based on conventional GDP measures, and in many cases would be negative. Many countries would find that their losses due to environmental degradation are greater than their gains due to increased production of goods and services.

If economic growth were redefined so as to take into account such losses, the only way then to achieve high rates of growth would be by increasing the production of goods and services without causing environmental destruction (or causing only small amounts of environmental destruction); alternatively, increases in the production of goods and services would include improvements in the quality of environmental resources resulting from investments in natural capital.

The problem of the conflict between growth (conventionally defined) and sustainability is related to economists' systematic neglect of the factor of production 'land'. Growth models in mainstream economics show how output increases in relation to labour and capital (in the sense of physical capital), sometimes also considering human capital, while completely ignoring land, which was taken to be an unimportant factor that was permanently fixed in quantity and quality. It is only in recent years that land has been redefined by environmental economists to consist of 'natural capital', which can be destroyed or improved through investments as discussed above in Section *Why economic growth occurs* (Supplementary material).

Thinking points

- How does the way we think of (or define) growth affect its relationship to the environment?
- Do we have a moral obligation to nature and the environment?
- Is our choice of how to measure growth based on economic or moral criteria (or both)?
- If governments began using alternative accounting methods, would production and consumption patterns necessarily change so as to become more consistent with sustainability?
- Some might argue that continuous growth is impossible because it is unsustainable to achieve indefinite growth. Is this argument valid?

Impact of economic growth on income distribution

A large number of studies have been carried out investigating the relationship between growth in GDP *per capita* and income distribution in developing and developed countries. The results have been inconclusive: while in some countries income distribution worsened in the early periods of growth and then improved, in some others the opposite happened, while in many others, income distribution did not show any clear pattern of change. These results lead to the conclusion that *there is no clear relationship between growth in GDP per capita and income distribution*; instead, what happens to income distribution as a country grows is a reflection of particular conditions in each country and the kinds of growth policies that are pursued.

For example, many countries in Latin America had highly unequal income distributions to begin with; income distribution in these countries has tended to remain highly unequal. A number of countries in East Asia (for example South Korea) had far more equal income distributions when they began their rapid growth, and this remained so even with rapid growth during the 1970s and 1980s. In addition, countries of East Asia placed a strong emphasis on the development of human capital, a policy that played a key role in ensuring broad-based participation in the benefits of growth, with positive effects on the equality of income distribution.

Yet, income inequalities in many countries around the world have been widening over the past three or so decades. They have been growing in China, India, Indonesia, Thailand and other East Asian and South-east Asian countries that had achieved greater income equality and reductions in poverty in their early years of growth. Russia and most other central and eastern European countries have similarly been experiencing sharp rises in income inequalities. A number of countries in Latin America have seen growing inequalities as well. Almost all OECD[4] countries also show worsening income distributions. Some of these countries have also experienced increases in the number of households below the poverty line.

In both developed and developing countries, a major factor behind increasing income inequalities has been the growing use of market-based supply-side policies (see Chapter 13). Economies in central and eastern Europe and the former Soviet Union that transitioned to market-based systems have additionally been influenced by the switch to market economies and the loss of government protection of vulnerable groups. In developing countries, income inequalities increased due to economic and trade liberalisation, which as we will see gives rise to both winners and losers (Chapter 20). While those who can take advantage of new opportunities gain, many become worse off, if they are less educated or skilled, cannot get credit, are geographically isolated, have nothing to produce for export, lose their jobs due to privatisations or reductions in the size of the government sector, and so on.

In addition, income distribution in developing countries can worsen as a result of economic growth due to inappropriate government policies, such as:

- the introduction of capital-using (labour-saving) technologies in industry and agriculture, which tend to use relatively more capital inputs in spite of relatively abundant supplies of labour, creating rural and urban unemployment (see Chapter 19)
- low levels of government investment in human capital, which negatively affect people on lower incomes and the poor disproportionately more than wealthier people
- allocating most services and infrastructure investments to urban areas and ignoring the rural sector where most of the poor live
- within the urban sector, concentrating infrastructure and services investments within the formal (modern and highly paid) sector and ignoring the urban slums.

It can therefore be concluded that *economic growth is neither 'good' nor 'bad' for income distribution; this instead depends very much on the kinds of policies countries adopt in order to achieve growth*.

Impact of economic growth on unemployment and inflation (Supplementary material)

If you are interested in this topic you may read about it in the digital coursebook as Supplementary material.

A concluding note

We have examined the effects of economic growth on several factors. Note that these factors can also impact on economic growth:

- **Living standards.** Economic growth can be expected to impact on living standards, but improved living standards measured as improvements in human capital or reduced income inequalities are major factors contributing to economic growth (see this chapter on causes of growth and Chapter 19).
- **Environment.** Economic growth that ignores the effects on the environment leads to environmental unsustainability, but unsustainability also leads to lower economic growth due to destruction of common pool resources. On the other hand, economic growth based on the principle of sustainable development leads to environmental preservation, which in turn can be expected to lead to higher economic growth in the future.
- **Distribution of income.** Economic growth can make the distribution of income more or less equal (equitable), but a more equal distribution of income has a positive effect on growth (see Chapter 19).

The likelihood of a two-way causality, where economic growth impacts upon factors such as the above, and where these factors in turn impact upon economic growth, sometimes makes it difficult in the real world to determine what causes what.

[4] The OECD is the Organisation for Economic Co-operation and Development, consisting mostly of developed countries.

TEST YOUR UNDERSTANDING 11.4

Discuss the possible negative and positive consequences of economic growth on

- a living standards,
- b the environment, and
- c income distribution. To what extent can governments make the consequences lean more towards the positive?

11.2 Sustainable level of government debt (HL only)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the relationship between a budget deficit and government (national) debt (AO2)
- explain that government (national) debt is measured as a percentage of GDP (AO2)
- explain the costs of a high government (national) debt on the following (AO2)
    - debt servicing costs
    - credit ratings
    - future taxation and government spending

The meaning of government debt

Government debt, also known as **national debt**, or *public debt*, refers to the amount of money that a government owes to lenders outside of the government itself.

How government debt arises

A *government budget* is a type of plan of a country's revenues and expenditures over a period of time (usually a year). Most of the government's revenues come from taxes. Its expenditures consist of spending on numerous items such as wages of government employees, provision of merit goods, investments in infrastructure like roads, and transfer payments to vulnerable groups like unemployment benefits and child subsidies. (We will examine government revenues and expenditures in more detail in Chapter 13.)

If tax revenues are equal to government expenditures over that period, the government is said to have a *balanced budget*. However, in practice, the government's budget is rarely if ever balanced. If expenditures are larger than tax revenues, there is a **budget deficit**; if expenditures are smaller than tax revenues, there is a **budget surplus**. When there is a budget deficit, the government finances (pays for) the extra expenditures over revenues by borrowing. This is similar to personal finance: if you spend more than you earn, it is likely that you borrow to pay for your extra spending over your income.

Governments very often have deficits. They have many commitments in terms of their spending, as they must provide health care, education, infrastructure, defense, and they pay salaries to their employees and make transfer payments. Often their commitments are greater than their revenues, and borrowing allows them to continue to spend without having to increase taxes. There is a greater need for borrowing during a recession because as unemployment increases, tax revenues fall while government spending on unemployment benefits rises (see Chapter 10).

Over time, the government's accumulation of deficits minus surpluses is referred to as *government debt*, or *national debt* or *public debt*. In any particular year, if the government runs a budget deficit, its debt will become larger; if it runs a budget surplus, its debt will become smaller. **Sustainable debt** refers to a level of debt where the borrowing government has enough revenues to meets it debt obligations (payment of interest and repayment of the borrowed amount) without accumulating arrears (overdue debt payments) while also allowing economic growth to continue at an acceptable level.

How governments borrow

Governments very commonly borrow by issuing bonds, which are a form of debt. When the government borrows to finance a deficit, it issues a certificate called a bond that promises to pay interest at various intervals until a certain date when the money is repaid to the bond holder. The holder of the bond is therefore the lender, and the issuer of the bond is the borrower. Financial investors, who may individuals, firms, banks, or any type of organisation, have the incentive to buy bonds because of the interest income they receive. Sometimes countries may also borrow directly from financial institutions (to be discussed in Chapter 20). The borrowing may be from internal sources, within the country, or from external sources, from other countries.

Measurement of debt as a share of GDP

One of the most common ways to measure the size of a country's government debt is as a share of GDP of the borrowing country. This is referred to as the *debt-to-GDP ratio*.

Table 11.4 shows the levels of debt as a share of GDP of selected countries. The country with the largest government debt as a share of GDP is Japan, followed by Greece and Italy.

Government debt has been increasing rapidly since the global financial crisis that began in 2008, with the largest increases generally occurring in developed countries.

Costs of high levels of debt

High levels of debt have a number of disadvantages for the economies of debtor countries.

Debt servicing costs

Debt servicing refers to the payments that must be made in order to repay the principal (the amount of the loan) plus interest payments. Large debt service payments have major opportunity costs because the government has fewer resources to spend on social services (health, education, etc.) and infrastructure.

In addition, the portion of the debt that is from external (foreign) lenders must be repaid in foreign exchange (foreign currencies). This means that the government is forced to use export earnings for debt servicing, resulting in less foreign exchange to pay for imports of needed capital equipment, other production inputs and goods and services generally. The foregone imports are an additional opportunity cost with negative consequences for economic growth. (This will become clearer to you have studied Chapter 16.)

Poor credit ratings

A credit rating is an assessment of the ability of a borrower to pay back loans, usually carried out by agencies that are qualified to do this. (Examples of such

Country	Debt-to-GDP ratio % 2018	Country	Debt-to-GDP ratio % 2018
Japan	234.2	Zimbabwe	72.6
Greece	181.8	Angola	71.6
Italy	127.5	Zambia	68.0
United States	109.5	India	67.3
Gambia	105.2	Sierra Leone	64.0
France	96.2	China	54.4
Brazil	90.2	Cameroon	34.1
Canada	83.8	Chile	24.6
United Kingdom	85.9	Bulgaria	22.8

Source: *http://worldpopulationreview.com/countries/countries-by-national-debt*

Table 11.4: Debt-to-GDP ratios in selected countries

agencies include Standard & Poor's, Fitch Ratings, and Moody's.) A high credit rating received by a government means that it is expected to be able to pay back its loans in full and on time without difficulties. A low credit rating means that it is expected that the government may have difficulties servicing its debt. This makes it more difficult for the borrowing government to find financial investors willing to lend (by buying the government's bonds) as well as more difficult to borrow from financial institutions. If a government has a high level of debt, shown by a high debt-to-GDP ratio, it is likely to receive a lower credit rating, creating difficulties for the government's ability to borrow in the future. This often forces the borrowing government to offer higher interest rates to financial investors in order to induce them to buy the bonds (in other words to lend), which increases the debt servicing costs to the government.

Impacts on future taxation and government spending

If a government wants to decrease the size of its debt, it must have budget surpluses rather than budget deficits. A budget surplus as we have seen above means that government revenues are greater than government spending. This extra amount of revenue over and above spending can be used to pay back a portion of the debt, which will work to reduce its overall size.

However, this may create serious difficulties for the government. In order to achieve budget surpluses, it must either increase taxes, or it must decrease spending. Both of these are politically unpopular. But more serious than the political consequences are the economic consequences of increased taxes or lower government spending. As you may remember from Chapter 9, increased taxes on consumer incomes reduce consumption spending (*C*) while increased taxes on business profits reduce investment spending (*I*). Reductions in both *C* and *I* cause aggregate demand to fall, resulting in lower real GDP. At the same time, decreases in government spending also cause aggregate demand to fall, exerting a further downward push on real GDP. The result is that as the government tries to achieve a budget surplus, it causes real GDP to fall, creating a recession or a deflationary gap (negative growth).

What happens then to the debt-to-GDP ratio? It actually increases! The government can then end up being worse off in terms of the size of its debt relative to GDP.

But the story does not end there. Once the recession begins, cyclical unemployment increases, and incomes fall, which mean that the government's tax revenues fall. At the same time, the government's spending on unemployment benefits increase. The fall in tax revenues and increase in government expenditures happen at the same time that the government is trying to increase its tax revenues and reduce its expenditures. As a result the government is confronted by a situation where to achieve a budget surplus it must increase taxes even more and cut spending even more, creating a vicious circle of further decreases in aggregate demand and falling real GDP.

This is what happened in the case of Greece, which as Table 11.4 shows has the second highest debt-to-GDP ratio in the world (see Real world focus 11.1 and 17.1, Chapter 17).

In fact, a country's debt-to-GDP ratio can be reduced in another, far more logical way. Very simply, this can be done through economic growth which as we know involves higher real GDP. As GDP increases, the debt-to-GDP ratio falls. For example, the United States after the Second World War had the very high debt-to-GDP ratio of 122%. Within ten years this had been cut in half due to economic growth, without the government having 'paid back' its debt.

Increased income inequality

Government debt is likely to increase inequality in income distribution. Buyers of government bonds, who are the lenders to the government, tend to be higher income people. When the government pays them interest, it does so through tax revenues. Therefore there is a transfer of income away from lower income taxpayers and toward higher income bond holders.

Lower private investment

Fears that a government may be unable to service its debts create uncertainty regarding economic conditions and scares away private investors, both domestic and foreign. Even if investment does take place, it is more likely to involve short-term investment projects with quick returns, rather than longer-term ones with greater potentials to support economic growth.

Possibility of a debt trap

As levels of debt rise, there comes a point where the level of debt cannot be sustained: new debt requires higher debt service payments, which require more foreign borrowing, which leads to more debt servicing payments, and so on, in a self-reinforcing spiral in which the country is trapped. This has been termed the 'debt trap', involving a situation where a country must keep on taking out new loans in order to pay back the old ones. Many countries, particularly in Latin America and sub-Saharan Africa, were caught in a debt trap during the 1980s and Greece has been caught in one in more recent years.

Lower economic growth

The above factors may work to lower economic growth in countries with high levels of debt, due to lower government spending, increased taxes, reduced investment and fewer imports of capital goods.

TEST YOUR UNDERSTANDING 11.5

1. Explain the meaning of government debt in relation to budget surpluses and deficits.
2. Discuss some of the costs of a high level of government debt.

REAL WORLD FOCUS 11.1

Austerity in the euro zone

The policy of increasing taxes and reducing government spending to deal with budget deficits and government debt is known as *austerity*. As we will discover in Chapter 13 it is well known that austerity is likely to lead to recession. However, in spite of that there is a major debate among economists about whether this is the appropriate policy to deal with deficits and debt.

When the financial crisis of 2008 occurred, many euro zone countries feared a debt crisis, in which their spending to pay back debt would be too high, in other words they feared their debt would become unsustainable. Greece in fact did face a serious debt crisis (see Real world focus 17.1, Chapter 17). They therefore began a policy of austerity.

Figure 11.5: Madrid, Spain. Protestors in the 'March for dignity' in 2014, protesting against the government's austerity programme and the social and economic crisis

Yet according to numerous studies, including two by The Institute of International Finance and Oxford Economics, the austerity policy led to lower GDP growth in the years following the crisis. This involved reductions of both actual GDP as well as potential GDP. In terms of the business cycle diagram in Figure 8.4 (Chapter 8) this has meant smaller expansions of actual output and a flatter long-term growth trend showing potential GDP. 'It was like letting air out of a bouncing ball. Its ability to bounce back got worse the more it was deflated.'[5] It is estimated that in the ten years since 2008, Europe lost an economy the size of Spain, with a GDP of $1.3 trillion and 19 million employed people.

Moreover, the debt issue was not solved. Instead, through lower real GDP growth, the debt burden was increased rather than reduced. According to the Harvard Business Review

> *Eurozone governments – especially those in struggling Southern European countries (Spain, Greece, or Portugal) – switched dramatically towards austerity in the years 2010-2014. Most experts now agree that these policies had such damaging and persistent negative effects on growth that they were self-defeating. Governments were reducing spending in order to bring their debt levels under control. But GDP fell so much that the actual effect was to push up the ratio of debt to GDP. As a result, debt became even less sustainable than before the austerity measures were implemented.*

Sources: *https://www.businessinsider.com*
https://hbr.org

[5] https://www.businessinsider.com/europe-austerity-is-decision-to-go-into-recession-2019-1

CONTINUED

Applying your skills

1 Use a business cycle diagram to show how austerity affected

 a short-term fluctuations, and

 b the long-term growth trend of euro zone economies.

2 Although a goal of austerity is to reduce the debt-to-GDP ratio, explain how this policy had the opposite effect in the euro zone.

11.3 Potential conflict between macroeconomic objectives

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- discuss the trade-off between inflation and unemployment based on: (HL only)
  - the short-run and long-run Phillips curves (AO3)
  - *AD-AS* diagrams (AO4)
  - Phillips curve diagrams (AO4)
- discuss the potential conflict between (AO3)
  - low unemployment and low inflation
  - high economic growth and low inflation
  - high economic growth and environmental sustainability
  - high economic growth and equity in income distribution

Low unemployment and low inflation

The trade-off based on the Phillips curve (HL only) and the potential conflict between low unemployment and low inflation were discussed in Chapter 10, Section *The conflict between low unemployment and low inflation*. The discussion came at the end of the discussion of low unemployment and low inflation which were the two macroeconomic objectives discussed in Chapter 10.

High economic growth and low inflation

This topic was introduced above under *Consequences of economic growth*. We can now discuss this further by referring to the distinction between demand-pull and cost-push inflation.

Demand-pull inflation and economic growth

Demand-pull inflation is caused by increases in aggregate demand. This can be shown both by use of the monetarist-new classical model and the Keynesian model as a rightward shift in the *AD* curve, shown in Figure 10.5 (Chapter 10).

In the Keynesian model, as long as *AD* increases along the horizontal portion of the *AS* curve, there is economic growth with no inflation. Therefore as long as the economy is operating in a deflationary gap, below potential output, there is no conflict between low inflation and economic growth as growth can occur with no inflationary pressures. The same is not true in the monetarist/new classical model, as here when the economy is in a deflationary gap, an increase in aggregate demand will result in both economic growth and an increase in the price level, suggesting a possible conflict.

However, as the economy approaches potential output, inflationary pressures appear also in the Keynesian model due to resource bottlenecks, suggesting the emergence of a conflict between economic growth and low inflation in this model as well.

The only way that further increases in aggregate demand will not be inflationary, in the context of both models, is if at the same time that aggregate demand is increasing there is an increase in long-run aggregate supply (*LRAS*) or Keynesian *AS*, shown by rightward shifts in these two curves as in Figure 11.3. We can see in this figure that as *AD* increases by the same amount as the *AS* curves, increases in real GDP are not accompanied by a higher price level. The reason is that as the economy's total demand for real GDP increases, there is a corresponding increase in the economy's ability to supply that real GDP. But if *AD* increases faster than *LRAS* or Keynesian *AS*, then increases in real GDP or economic growth will result in inflation.

Cost-push inflation

This is caused by decreases in short-run aggregate supply due to such factors as higher prices of factors of production. As Figure 10.6 (Chapter 10) shows, the leftward shift in the *SRAS* curve leads to a higher price level and a fall in real GDP, or negative economic growth, also known as s*tagflation*. Therefore, with cost-push inflation it is not possible to have positive economic growth at the same time as the price level is rising.

High economic growth and environmental sustainability

This topic was explored above under *Consequences of economic growth*, where it was concluded that economic growth and environmental sustainability can be pursued together provided governments take appropriate measures to ensure sustainable resource use. This is the meaning behind the concept of *sustainable development*.

High economic growth and equity in income distribution

This topic was also explored under *Consequences of economic growth*. It was concluded that economic growth is neither 'good' nor 'bad' for income distribution; this instead depends very much on the kinds of policies countries adopt in order to achieve and handle growth.

TEST YOUR UNDERSTANDING 11.6

Using your knowledge of economics and the information provided in this chapter and Chapter 10, discuss the extent to which there may be a conflict between:

a low unemployment and low inflation

b high economic growth and low inflation

c high economic growth and environmental sustainability

d high economic growth and equity in income distribution.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Select a country you are interested in. Examine possible conflicts between its growth and environmental sustainability. Investigate whether any measures are being taken to deal with the environment.

2 There are some developed countries with high debt-to-GDP ratios. Research such a country of your choice. Does a large ratio always mean the country is struggling? Why can some developed countries (such as the United States) maintain high debt-to-GDP ratios while others suffer immensely?

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 12

Economics of inequality and poverty

BEFORE YOU START

- In most societies, we see that there are some people who have high incomes and accumulate a lot of wealth while others have very little. What do you think might be some reasons for such extreme differences in income and wealth?
- What, if anything, do you think governments should do to reduce extreme inequalities in income and wealth?

Poverty and inequalities in income and wealth are major issues in countries around the world. While poverty is more prominent in developing countries, it is present in rich countries as well. This chapter will discuss causes and consequences of poverty and inequality, their measurement and policies to tackle them.

Photo: Morocco. Division between rich and poor

12.1 Inequality

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the relationship between equity and equality (AO2)
- explain economic inequality as (AO2)
  - unequal distribution of income
  - unequal distribution of wealth
- use the Lorenz curve and Gini coefficient (index) to measure economic inequality (AO2)
- draw Lorenz curves to show the distribution of income and changes in the distribution of income (redistribution) (AO4)
- construct a Lorenz curve based on income quintile data (HL only) (AO4)

The relationship between equity and equality

We encountered the concepts of equity and equality in Chapter 1 (see Section 1.4). As *equity* involves fairness, something is *equitable* if it is fair. This is a normative concept because different people have different ideas and beliefs about what is fair. On the other hand *equality* or the idea of being the same is a positive concept because something may be equal or unequal on the basis of some measure.

The concepts of equity and equality are used in economics mainly to describe the distribution of income and/or wealth. As we know from our study of the circular flow model (Chapters 1 and 8), income is the money received by owners of factors of production. *Wealth*, on the other hand, refers to the money, assets or things of value that people own, such as savings deposits (money saved in a bank); stocks in the stock market; bonds; land, houses and other property; valuable paintings or jewellry, and so on.

While equity differs from equality, it is usually interpreted to mean equality. Therefore the expressions a 'more equitable' or 'more equal' distribution of income or wealth usually mean the same thing. Both expressions are correct, provided it is understood that in these cases, *equity is interpreted as greater equality (or less inequality)*. The reason that equity is most often interpreted in this way is that there is a widely shared belief around the world that highly unequal distributions of income are unfair. Therefore governments around the world usually have in place policies that try to reduce income inequalities.

Economic inequality

The meaning of economic inequality

Economic inequality refers to the degree that people in a population differ in their ability to satisfy their economic needs; it means inequality in living conditions that arise due to monetary factors. There are many sources of economic inequality, including income and wealth, education, health, nutrition, gender and more, but economists focus on inequalities that result mainly from *differences in income and wealth*.

Unequal distribution of income

Income inequality arises from differences in how evenly income is distributed in a population. Income includes the money that people receive from their employment as well as other sources including interest from savings accounts and holdings of bonds, dividends from shares (stocks) in the stock market, rents from property that is owned and rented out, pensions or government benefits.

Unequal distribution of wealth

Wealth inequality arises from differences in the amount of wealth people own, which as noted above refers to the money or things of monetary value including savings, stocks, land, houses, and more.

Both income and wealth are generally distributed unequally, so that some groups have much more income and/or wealth than other groups. This applies both to populations within countries, as well as across countries.

How economic inequality is measured

Table 12.1 presents data on income distribution of selected countries around the world. The data show how income is distributed by *quintiles* of the population. A **quintile** is a 20% portion of a country's population; we can divide a population into five quintiles, ranging from the lowest (the poorest 20% of the population) to the highest (the richest 20%). If income were completely equally distributed, everyone would receive exactly the same income, so every quintile would receive 20% of income. However, in the real world this is a virtual impossibility. In all countries in the world, the presence of inequalities in income distribution means that the poorest quintile of the population receives less than 20% of income, and the richest quintile more than 20%.

This can be seen in Table 12.1. For example, in Brazil the lowest quintile receives 3.2% of income and the highest quintile 57.8%. In Belarus, the lowest quintile receives 9.9% and the highest quintile 35.5%. The higher the percentage share of income received by the poorest quintile, and the lower the percentage share received by the highest quintile, the more equal the distribution of income. Therefore, income distribution in Belarus is more equal than in Brazil.

Income shares can also be shown by *deciles*, which are 10% portions of the population (there are ten deciles) as well as *quartiles*, or 25% portions of the population (there are four quartiles). Sometimes income shares are broken down into 1% particularly for the top, or even the top 0.1%.

The Lorenz curve

A **Lorenz curve** is used to show the degree of income inequality in an economy. Named after an American economist Max Otto Lorenz, who devised this measure of income inequality in 1905, it is a visual representation of the kind of income distribution data in Table 12.1. To construct a Lorenz curve, we draw a square box, as in Figure 12.1 where the vertical axis measures the total amount of income in an economy in cumulative percentages (therefore it runs from 0 to 100%), and the horizontal axis plots the total population in the economy, also in cumulative percentages (from 0 to 100%). ('Cumulative' means that 20 represents the poorest 20% of the population, 40 represents the poorest 40%, and so on.) The diagonal line in the diagram represents perfect equality, as it shows that if income were perfectly equally distributed, 20% of the population would receive 20% of income, 40% would receive 40% of income, and so on. The Lorenz curve plots the *actual relationship* between percentages of the population and the shares of income they receive.

Country	Poorest 20%	Second 20%	Third 20%	Fourth 20%	Richest 20%	Gini coefficient
Australia 2014	6.8	12.0	16.1	22.1	43.0	0.36
Belarus 2017	9.9	14.2	17.9	22.5	35.5	0.25
Bolivia 2017	4.1	9.5	15.0	22.8	48.5	0.44
Brazil 2017	3.2	7.4	12.2	19.5	57.8	0.53
China 2015	6.4	10.6	15.3	22.3	45.4	0.39
Denmark 2015	9.4	13.9	17.2	21.8	37.7	0.28
S. Africa 2014	2.4	4.8	8.2	16.5	68.2	0.63
United Kingdom 2015	7.5	12.2	16.8	23.0	40.6	0.33
United States 2016	5.0	10.2	15.3	22.6	46.9	0.42

Source: *https://databank.worldbank.org/data/source/poverty-and-equity*

Table 12.1: Distribution of income by quintiles and gini coefficients in selected countries

In general, the closer a Lorenz curve is to the diagonal representing perfect income equality, the greater is the equality in income distribution. As we can see in Figure 12.1, Belarus clearly has greater income equality than Brazil.

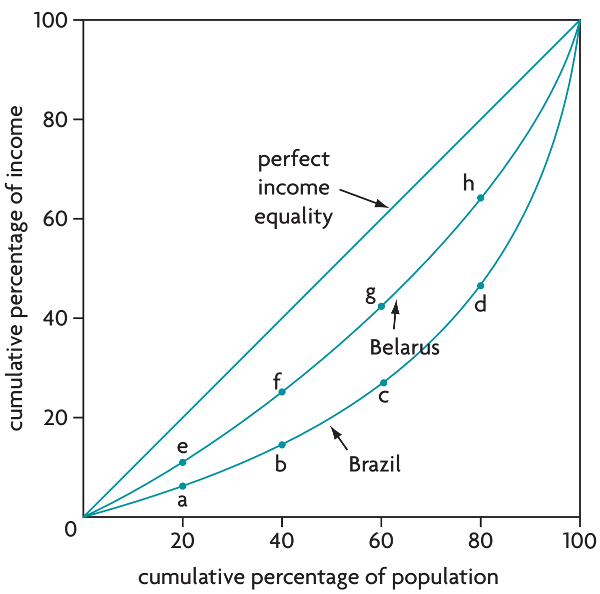


Figure 12.1: Lorenz curves: Belarus achieves greater income equality than Brazil

How to construct a Lorenz curve from income quintile data (HL only)

Figure 12.1 plots two Lorenz curves, one for Brazil, and one for Belarus (based on the data in Table 12.1). In the case of Brazil, the poorest 20% of the population receive 3.2% of income; this is shown by point a. Point b on Brazil's curve is obtained by adding the 3.2% of income of the poorest quintile to the 7.4% of income received by the second quintile, giving 10.6%, or the cumulative income of the bottom 40% of the population. Similarly, point c is obtained by adding the percentages of income received by the bottom three quintiles, giving 22.8% of income, and finally to find point d we add the incomes of the bottom four quintiles, getting 42.3% of income for 80% of the population. When these points are joined together starting from 0 and going up to 100% of the population, we obtain Brazil's Lorenz curve. Points e, f, g and h on Belarus' curve are calculated and plotted in exactly the same way. As expected Brazil's Lorenz curve is further away from the line of perfect equality indicating greater income inequality.

Note that to plot a Lorenz curve, we could use income distribution figures that divide the population into ten deciles (or tenths), or any other convenient subdivision.

The Gini coefficient

The **Gini coefficient** (or **Gini index**), named after Corrado Gini, an Italian statistician, is a summary measure of the information contained in the Lorenz curve of an economy. It is defined as

$$\text{Gini coefficient} = \frac{\text{area between diagonal and Lorenz curve}}{\text{entire area under diagonal}} = \frac{A}{A+B}$$

Where A and B represent the areas shown in Figure 12.2.

The Gini coefficient has a value between 0 and 1. If there were perfect income equality, the coefficient would be zero, since the numerator of the ratio would be zero. The larger the Gini coefficient, and the closer it is to 1, the greater is the income inequality, since the further away is the Lorenz curve from the diagonal. (A perfectly unequal income distribution would be where a single household receives all the income of the economy, and the numerator would be equal to the entire area under the diagonal, making the Gini coefficient equal to 1.)

Note that some publications express Gini coefficients as a percentage rather than a decimal. For example, a coefficient of 0.27 would appear as 27.0. This does not

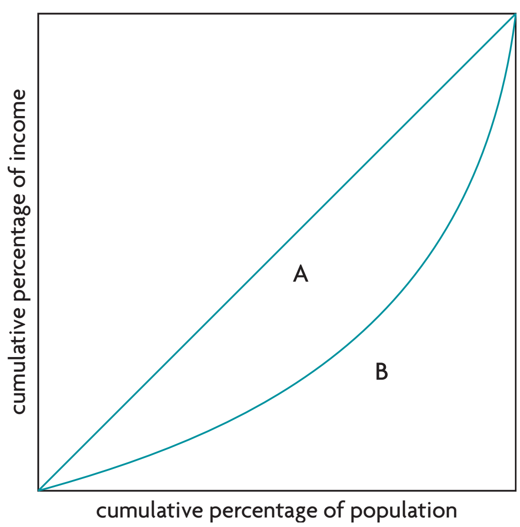


Figure 12.2: Deriving the Gini coefficient from a Lorenz curve

in any way change in the meaning of the Gini coefficient, which in this method of expression has a value ranging between 0 and 100.

The last column in Table 12.1 shows Gini coefficients that correspond to each of the income distributions. Belarus' Gini coefficient of 0.25 and Brazil's of 0.53 clearly indicate that Belarus has a relatively more equal income distribution.

> The *Gini coefficient* is a summary measure of income inequality. In a Lorenz diagram it is the ratio of the area between the diagonal and the *Lorenz curve*, to the total area under the diagonal. It has a value between 0 and 1; the closer the value is to 0, the greater the income equality; the closer the value is to 1, the greater the income inequality.

Wealth inequality

Everything that has been said above about measurement of income inequality applies also to wealth inequality. The three methods discussed above, namely (i) quintiles (or deciles or quartiles), (ii) Lorenz curves, and (iii) Gini coefficients can be used in exactly the same way to show the extent of wealth inequality.

The distribution of wealth is generally far more unequal than the distribution of income in most countries in the world. On average, Gini coefficients in the case of wealth distribution are roughly double the size of the Gini coefficients of income distribution. This is shown in Figure 12.3. We can see here that developed countries on the whole have slightly lower wealth and income inequality than emerging market economies, but in both groups wealth inequality is far greater than income inequality.

Reasons behind greater wealth inequality include the following:

- Limited growth in wages makes it difficult for low-income and middle-income people to save and accumulate wealth.
- High-income people tend to consume a smaller fraction of their income than lower-income people therefore have greater possibilities of saving and accumulating wealth.

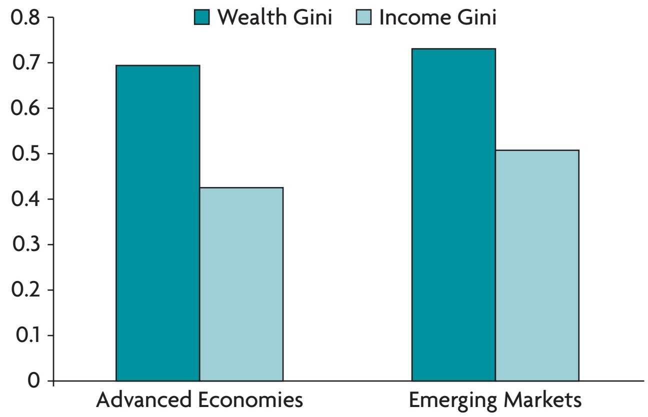


Figure 12.3: Income and wealth inequality in advanced economies* and emerging market economies**

* Advanced economies include developed countries

** Emerging market economies include Argentina, Brazil, China, India, Indonesia, Mexico, Pakistan, Thailand and Turkey

Source: *www.imf.org*

- Income and wealth inequalities feed on each other. The greater the income, the more possibilities for accumulating wealth, but many types of wealth (stocks, bonds, real estate) lead to even more income and hence even more possibilities for accumulating more wealth. For example, in the United States, in 2015 the share of income that came from wealth for the richest 1% of the population was nearly 60%, while the share of income coming from wealth for the *bottom half* (or 50%) of the population was about 5%.[1]

Using Lorenz curves to illustrate income and wealth redistribution

Later in this chapter, we will consider methods governments can use to redistribute income and wealth, to make their distribution more equal. Graphically, this appears as a shift of a country's Lorenz curve closer to the diagonal line, and is reflected in a lower Gini coefficient. Figure 12.4 shows how a Lorenz curve shifts towards the diagonal after the government pursues policies to redistribute income or wealth to reduce the degree of economic inequality in the economy.

[1] https://theweek.com/articles/717294/wealth-inequality-even-worse-than-income-inequality

Figure 12.4: Lorenz curves and income redistribution

TEST YOUR UNDERSTANDING 12.1

1 Outline the meaning of equity and equality and how they relate to each other.

2 Using examples, explain the meaning of economic inequality.

3 Outline the meaning of quintiles, and how we use them to measure income distribution.

4 List the countries in Table 12.1 in order of increasing income inequality and explain your reasoning.

5 a Define the Gini coefficient.

b Using a diagram, describe how we derive it from a Lorenz curve diagram.

c State its possible range of values.

d Outline how we interpret its possible values with regard to income equality/inequality.

6 a Choose any two countries from Table 12.1 and draw their Lorenz curves in a single diagram (HL students should plot the Lorenz curves).

b Outline how you use your Lorenz curve to determine which country has a more equal distribution of income.

c Describe how you can use the corresponding Gini coefficients to compare the income distributions of the two countries.

7 a Outline how redistribution of income changes the position of a country's Lorenz curve.

b Describe how redistribution would be reflected in a Gini coefficient. (Make sure you distinguish between redistribution that increases or decreases income equality.)

8 Describe the methods that can be used to measure wealth inequality.

REAL WORLD FOCUS 12.1

Trends in wealth concentration

In the United States, wealth concentration had peaked in 1929 when the top 0.1% of the population owned nearly 25% of wealth. It then began to fall steadily, reaching a low in 1978 when the top 0.1% owned 7% of wealth. Since then it has been steadily climbing again. In the period before 2020 the richest 0.1% owned over 20%, while the richest 1% owned 40% of wealth.[2] Moreover, the three richest Americans held more wealth than the bottom 50% of the population of the United States.[3]

Figure 12.5: Bel Air Los Angeles, USA. Neighborhood with mansions and golf course

In England (which is part of the United Kingdom) less than 1% of the population owns half of England's land. This includes about 25 000 landowners, consisting mainly of the aristocracy and corporations. By contrast, all of England's homeowners (in a population of about 55 million people) own 5% of the land.[4]

High wealth concentration exists on a global level as well, where wealth inequality is also growing. It is estimated that the richest 1% of the global population owns half of the world's wealth.[5] The top 10% of the combined populations of China, Europe and the US own more than 70% of wealth, while the bottom 50% owns less than 2% of wealth. If countries in Latin America, Africa and the rest of Asia were included, wealth concentration would be even greater as in most countries the share of wealth owned by the bottom 50% is close to zero.[6]

If trends in wealth accumulation that began in 2008 with the Global Financial Crisis continue, the world's richest 1% will control two-thirds of the world's wealth by 2030. Since 2008, the wealth of the richest 1% has been growing at about 6% per year, compared to 3% per year for the remaining 99%.[7]

Applying your skills

1. Draw Lorenz curves to illustrate the changes observed in wealth distribution in the United States from the early 20th century up to the present.
2. Draw Lorenz curves to illustrate the change in global wealth distribution since 2008.

[2] https://www.nber.org/papers/w25462.pdf

[3] https://www.forbes.com/sites/noahkirsch/2017/11/09/the-3-richest-americans-hold-more-wealth-than-bottom-50-of-country-study-finds/#272cc9e33cf8

[4] https://www.theguardian.com/money/2019/apr/17/who-owns-england-thousand-secret-landowners-author

[5] https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf

[6] https://www.nber.org/papers/w25462.pdf

[7] https://www.theguardian.com/business/2018/apr/07/global-inequality-tipping-point-2030

12.2 Poverty

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the difference between absolute and relative poverty (AO2)
- explain how poverty is measured using (AO2)
    - single indicators including
        - international poverty lines
        - minimum income standards
    - composite indicators including
        - the Multidimensional Poverty Index (MPI)
- explain difficulties in measuring poverty (AO2)

Understanding poverty and its measurement

The meaning of poverty

Poverty refers to an inability to satisfy minimum consumption needs. Beyond this general definition, there are two different perspectives on how best to define poverty: in an absolute sense and in a relative sense.

Absolute poverty and its measurement

Absolute poverty refers to a situation where a person or family does not have enough income to meet basic human needs. Measures of *absolute poverty* begin by defining a minimum income level called a **poverty line**. According to the OECD, a poverty line is: 'An income level that is considered minimally sufficient to sustain a family in terms of food, housing, clothing, medical needs and so on.'[8] Most countries have a national poverty line, determined by government authorities as an appropriate amount of income required to satisfy minimum needs. In addition, the World Bank (an international organisation that lends to developing countries for development purposes; see Chapter 20) has determined an international poverty line to be:

- living on less than $1.90 a day, which is defined as *extreme poverty*; this figure is periodically adjusted to take inflation into account

Once a poverty line has been determined, the amount of poverty is found by taking the percentage of a population (or the number of individuals) whose income falls below the poverty line.

Data on extreme poverty in selected countries around the world, compiled by the World Bank, appear in Table 12.2. Note that the table includes developed countries as well as developing ones.

In a major study on global poverty in 2018, the World Bank concluded that there has been significant progress in reducing extreme poverty, which fell from 36% of the global population in 1990 to 10% in 2015. According to the Bank, the greatest progress was made in East Asia and South Asia, while sub-Saharan Africa showed a much slower pace of poverty reduction.

Country	Poverty rate %	Country	Poverty rate %
Iceland	0	India	13.4
Portugal	0.1	Bangladesh	15.2
Norway	0.2	South Africa	18.9
Ireland	0.2	Ethiopia	27.0
United Kingdom	0.2	Chad	34.1
Australia	0.5	Tanzania	40.7
China	0.7	Rwanda	51.5
Brazil	3.4	Mozambique	62.2
Philippines	8.3	Central African Republic	77.7

Source: *https://openknowledge.worldbank.org/bitstream/handle/10986/30418/9781464813306.pdf*

Table 12.2: Extreme poverty (living on less than $1.90 per day) in selected countries, 2015

[8] OECD, Glossary of Statistical Terms, Poverty Line.

As noted above, in addition to the World Bank's *extreme poverty* line, both developed and developing countries also have national poverty lines. These are usually set at a higher income level than that of the World Bank, so that a larger percentage of people are considered poor by national standards.

The World Bank has been criticised on the grounds that its international poverty line of $1.90 per day is too low to be meaningful in terms of poverty reduction.[9] For example, in South Africa the Bank's international poverty line puts 18.9% of the population below the poverty line, while according to the South African government's poverty line 55% of the population are poor.[10]

In acknowledgement of this point the World Bank presents two more poverty lines:

- living on less than $3.20 a day is a poverty line for lower-middle-income countries
- living on less than $5.50 a day is a poverty line for upper-middle-income countries.[11]

Nearly half of the world's population lives below the poverty line of $5.50 per day.[12]

In addition, in recognition of the limitations of international poverty lines, the World Bank has proposed a *multidimensional* measure of poverty that takes into account dimensions of poverty in addition to income, to be discussed below.

Relative poverty and its measurement

Relative poverty is a concept that compares the income of individuals or households in a society with median[13] incomes. It is closely related to how equally or unequally society's income is distributed among its total population. If income were equally distributed, there would be no relative poverty, since no one would be poor relative to someone else. In general, the more unequal the distribution of income, the greater is the degree of relative poverty.

The idea behind relative poverty is that poverty is much more than being unable to afford a minimum of basic goods and services. Even though people may be able to buy basic necessities, they are still poor if they cannot afford goods and services and a lifestyle that are typical in a society. The measurement of what is 'typical' is based on a standard determined by the median income level. If people's incomes fall far below this median level, they are considered poor. Measurement of relative poverty involves specifying a particular percentage of median income below which there is poverty. Often this is taken to be 50%. For example, say the median annual family income in an economy is $20 000. Taking 50% of this, we have $10 000. Any family whose annual income falls below $10 000 is considered poor (in relative terms). Table 12.3 presents data on relative poverty in some developed countries.

Whereas the discussion above has been in terms of *national* poverty rates, it is important to note that poverty rates differ widely among social groups in a society. In general, older people, children, single-parent households, women, and racial and ethnic groups that suffer discrimination, face higher poverty rates than national averages. This applies to most countries in the

Country	% of population living below 50% of median income (2017 or latest available)	Country	% of population living below 50% of median income (2017 or latest available)
Iceland	5.4	Germany	10.4
Denmark	5.5	United Kingdom	11.1
Netherlands	8.3	Australia	12.1
Norway	8.4	Canada	12.4
Belgium	9.7	Greece	14.4
Ireland	9.8	Latvia	16.9
Hungary	10.1	United States	17.8

Source: *https://data.oecd.org/inequality/poverty-rate.htm*

Table 12.3: Relative poverty in economically more developed countries, 2017

9 https://www.brookings.edu/blog/future-development/2018/11/14/why-the-world-bank-is-taking-a-wide-angle-view-of-poverty

10 https://theconversation.com/why-the-world-banks-optimism-about-global-poverty-misses-the-point-104963

11 Lower-middle-income countries have a *per capita* GNI of $996–$3895 while upper-middle-income countries have a *per capita* GNI of $3896–$12 055 https://blogs.worldbank.org/opendata/new-country-classifications- income-level-2018-2019

12 https://www.worldbank.org/en/news/press-release/2018/10/17/nearly-half-the-world-lives-on-less-than-550-a-day

13 The 'median' is the number that is in the middle of a series of numbers. Therefore, the median income is that income that lies in the middle of all income levels, so that half of income levels are greater and half are lower. Note that the median is different from the average or mean.

world, both more and less developed. Table 12.4 shows how poverty rates vary in the United States by race and by age for each race.

	Under 18 years % in poverty	18–64 years % in poverty	65 years and over % in poverty
Black	28.8	18.3	19.3
Hispanic	24.7	15.0	17.0
Asian	11.1	9.5	10.8
White	10.5	8.6	7.0

Table 12.4: United States poverty rates by race and age, 2017[14]

While the discussion here is about inequality by social groups in poverty, it might be noted that there is inequality by social groups in wealth as well. There is a large racial gap in wealth in the United States, which is worsening as Black wealth is decreasing while White wealth is increasing. According to projections, in 2020, White households will own 86 times more wealth than Black households. If present trends continue, in 2053 Black median wealth will have hit zero, while White median wealth will increase to $137 000.[15] These developments indicate that Black households on average are at risk of becoming impoverished.

Both absolute and relative poverty measures, on a national level and for particular social groups, are very useful to governments as guides to policies providing income support (such as transfer payments; see Section 12.5) as well as measures intended to combat poverty.

Measuring poverty by use of minimum income standards

Minimum income standards (MIS) refer to a method to measure poverty developed by the Joseph Rowntree Foundation in the UK. The method consists of ongoing research on what people in a population believe are the essentials for a minimum acceptable standard of living that allows people to participate in society. The MIS then produces budgets for a basket of goods including numerous essential items like food, clothing, housing, childcare, fuel costs and social and cultural participation, required by households in order to achieve the minimum standard of living. Based on this information it calculates the minimum income that is required for different family types (according to number of people, ages, geographical areas, etc.) to be able to buy the essentials in the basket.

The MIS reveals important information about:

- the number of people living below the minimum income required to buy the essentials
- the relative contribution of each item in the basket to households' abilities to achieve MIS
- how these change over time.

This information can be helpful to government as a guide to making policies to deal with poverty.

Calculations of the MIS began in the UK in 2008. Several countries have been conducting pilot (experimental) studies with the MIS including France, Ireland, Japan, Mexico, Portugal, Singapore, South Africa and Thailand.

Composite indicators

Composite indicators are measures of complex phenomena that cannot easily be described by a single indicator. They therefore try to capture more than one dimension of the issue in question. We will study composite indicators in greater detail in Chapter 18.

Multidimensional Poverty Index (MPI)

The **Multidimensional Poverty Index (MPI)** was developed in 2010 by the United Nations Development Programme and the Oxford Poverty and Human Development Initiative. It measures poverty in three dimensions: health, education and living standards. Each of these dimensions is intended to reflect *deprivations* (essential things that people do not have) measured by the following ten indicators:

• Health is measured by • child mortality • nutrition • Education is measured by • years of schooling • school attendance	• Living standards are measured by • cooking fuel • sanitation • drinking water • electricity • housing • assets

[14] https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html
[15] https://ips-dc.org/report-the-road-to-zero-wealth

THEORY OF KNOWLEDGE 12.1

Absolute and relative poverty

The concepts of absolute and relative poverty have different implications for the meaning of poverty. The concept of absolute poverty, based on an absolute poverty line that does not change over time (except for adjustments for inflation), suggests that with economic growth everyone will eventually rise above the poverty line. Therefore, many countries that experience long-term growth have been seeing falling absolute poverty rates. The relative poverty line, by contrast, changes constantly over time as incomes grow, and the poor are those who cannot keep up with rising average/median incomes. Relative poverty has actually been *increasing* in many countries over recent decades, due to *increasing* income inequalities.

The following famous passage, written by Adam Smith (the 'Father of Economics') in the 18th century in his classic book, *The Wealth of Nations*, offers an explanation of the difference between the two meanings of poverty, by referring to the meaning he attaches to 'necessaries':

> *'By necessaries I understand, not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without. A linen shirt, for example, is, strictly speaking, not a necessary of life. The Greeks and Romans lived, I suppose, very comfortably, though they had no linen. But in the present times, through the greater part of Europe, a creditable day-labourer would be ashamed to appear in public without a linen shirt, the want of which would be supposed to denote that disgraceful degree of poverty, which, it is presumed, nobody can well fall into without extreme bad conduct . . .* Under necessaries therefore, I comprehend, not only those things which nature, but those things which established rules of decency have rendered necessary to the lowest rank of people.'[16]

According to Adam Smith, a person who has the basic necessities of life required for physical survival but does not have a linen shirt is relatively poor, but not absolutely poor. One who does not even have the basic necessities of life is absolutely poor (and, of course, also relatively poor).

As noted in the text, measures of poverty are important because they form the basis for anti-poverty programmes such as transfer payments. Different countries use different measures for this purpose. For example, in the United States the official poverty measure used to determine eligibility for government assistance is an absolute one; in the European Union, the official poverty measure is a relative one (though both calculate absolute and relative poverty rates).

Thinking points

- Does society have a moral obligation to help the poor?
- Adam Smith identifies 'necessaries' to be 'those things which established rules of decency have rendered necessary'. Would he define poverty in the absolute or in the relative sense?
- What kinds of criteria are important for making a choice between absolute and relative poverty as the basis for anti-poverty programmes (social scientific, ethical or other)?
- A society's choice between an absolute or relative poverty measure as the basis for policy rests on some principle of equity. What do you think might be an equity principle for the US use of absolute poverty and for the EU use of relative poverty? How do the equity principles differ from each other?

[16] Adam Smith (1937) *An Enquiry into the Nature and Causes of the Wealth of Nations*, New York, Modern Library, pp. 821–2 (Book V, Chapter II, Part II, Article 4th).

As this list of indicators shows, the MPI goes beyond considering poverty solely in terms of *incomes*, considering instead a variety of areas in which poor people experience *deprivations*. The emphasis on deprivations was introduced by the Nobel Prize winning Indian economist Amartya Sen (see Chapter 18).

The MPI is used as a measure of poverty in developing countries (these will be studied in Chapters 18–20). It includes 105 countries (as of 2019) covering 5.7 billion people or 77% of the world's population.

Each country receives an MPI value from 0 to 1, where the higher the MPI value the greater the poverty. In order to be counted as *poor*, people must be deprived in at least one-third of the indicators listed above. Information is provided on the number of people and the percentage of the total population who are poor in each country. An important advantage of the MPI is that it can be broken down by indicator so that for each country it is possible to determine which indicators make the most important contributions to poverty.

For example, Tajikistan and Peru have similar MPIs, which are 0.049 for Tajikistan and 0.052 for Peru. But in Peru 18% of poverty is due to deprivations in years of schooling while in Tajikistan only 1% of poverty is due to this indicator. By contrast, in Tajikistan 35% of poverty comes from malnutrition while in Peru it is about 18%.[17]

Table 12.5 shows that nearly one-quarter of the population of the 105 countries live in multidimensional poverty. Most of the world's MPI poor, or 89%, live in South Asia or sub-Saharan Africa.

Developing region	MPI	Number of poor people (millions)	% of poor in total population*
Eastern Europe and Central Asia	0.009	3.5	2.4%
East Asia and the Pacific	0.025	117.7	5.9%
Latin America and the Caribbean	0.033	30.7	7.7%
Arab States	0.098	65.7	19.2%
South Asia	0.143	545.9	31.3%
Sub-Saharan Africa	0.317	559.4	57.7%
Global MPI Developing regions	0.115	1.33 billion	23.2%

* Poor people are defined to those who are deprived in one-third or more of the indicators

Source: *https://ophi.org.uk/wp-content/uploads/G-MPI_2018_2ed_web.pdf*

Table 12.5: Multidimensional poverty by region 2018

While the MPI considers poverty in developing countries, as we have seen there is poverty in developed countries as well. According to the United Nations Development Programme:

> '*Deprivations are a universal problem afflicting people in developed and developing countries alike. An average of 11 percent of the population in Organisation for Economic Co-operation and Development (OECD*[18]*) countries were below the income poverty line in 2014. As of 2012 there were 633 000 homeless people in the United States and 284 000 in Germany . . . An average of 15 percent of young people ages 15–29 are neither employed nor in education or training and are struggling to find their place in society. Health deprivations caused by obesity are also high. The most recent survey data indicate that an average of 53.8 percent of the adult population in OECD countries is overweight or obese and faces high risks of cardiovascular disease, respiratory illnesses, diabetes and other diseases.*'[19]

[17] https://ophi.org.uk/wp-content/uploads/G-MPI_2018_2ed_web.pdf

[18] The OECD includes 36 countries most of which are developed.

Multidimensional Poverty Index (MPI) of The World Bank

The World Bank (see Chapter 20) is in the process of developing another MPI. As noted by the Bank, the standard monetary measure of poverty does not capture important aspects of well-being, such as access to health care or a secure community. It therefore proposes a new MPI to complement that of the UNDP and Oxford by including a monetary indicator (income *per capita*) as well as some additional indicators. The World Bank's proposed indicators are shown in Table 12.6.

Income per capita	Electricity
Child school enrolment	Coverage of key health services
Adult school attainment	Malnourishment (child and adult)
Limited-standard drinking water	Incidence of crime
Limited-standard sanitation	Incidence of natural disaster

Source: *https://openknowledge.worldbank.org/*

Table 12.6: World bank proposed indicators for Multidimensional Poverty Index

Difficulties in measuring poverty

The measurement of poverty is a challenging task for several reasons.

Poverty has different meanings and different approaches to measurement

As we have seen there are different measures of poverty, depending on how this is interpreted. There is the distinction between absolute poverty and relative poverty, which gives rise to very different estimates on the extent of poverty. In addition, there is poverty measured on the basis of income, as well as poverty measured on the basis of deprivations in a number of different non-monetary areas, known as multidimensional poverty. All these are not consistent with each other.

Measurement problems

- Often, as we have seen, poverty is measured on the basis of income of a household. However, people also have some wealth to lesser or greater degrees, or they may have some savings, on which they can fall back in hard times. Income measures of poverty do not take wealth or savings into consideration.
- In some cases poverty is measured by use of household surveys. This raises several issues:
  - the information provided by the households that are surveyed is subjective, so that different people may have different opinions about their economic situation
  - such surveys do not include homeless people and people in institutions who are much more affected by poverty, resulting in underestimates of the extent of poverty
  - income figures may be understated in cases where there is freelance work or income from investments, resulting in overestimates of poverty.
- Urban areas usually have a higher cost of living than rural areas, so national poverty lines often exclude many poor in urban areas who cannot afford necessities like food, housing.
- Poverty lines tell us how many people (or the percentage of people) fall below the poverty line, but do not provide any information on how much they fall below the poverty line. In one case the majority of poor may be below but close to the poverty line, whereas in another they may be far below. Clearly there is greater poverty in the second case than in the first.

Overestimation or underestimation of the national poverty line

Depending on particular goals of governments, the national poverty line (for absolute poverty) may be overestimated or underestimated.

Overestimation results in a larger proportion of a population whose income falls below this line. This can be used by governments to argue in favour of receiving more foreign aid or multilateral assistance (see Chapter 20).

Underestimation results in a smaller proportion of a population with an income below this minimum. This can be used by governments formulating national strategies for poverty reduction, in the event that they would like to spend less on poverty reduction than on other activities demanding government funding.

[19] http://hdr.undp.org/sites/default/files/2016_human_development_report.pdf

TEST YOUR UNDERSTANDING 12.2

1 a Explain the difference between absolute and relative poverty.

b Outline whether it makes sense to compare poverty rates measuring absolute poverty with rates measuring relative poverty.

2 Distinguish between poverty lines and minimum income standards as measures of poverty.

3 The World Bank lends to developing countries for development purposes. Outline whether you think it makes sense to use the World Bank's definitions of absolute poverty to measure poverty in developed countries.

4 Using examples, explain possible advantages of composite indicators over single indicators as measures of poverty.

5 Discuss difficulties surrounding efforts to measure poverty.

12.3 Causes of economic inequality and poverty

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain causes of economic inequality and poverty including (AO2)
  - inequality of opportunity
  - different levels of resource ownership

CONTINUED

  - different levels of human capital
  - discrimination (race, gender and others)
  - unequal status and power
  - government tax and benefits policies
  - globalisation
  - technological change
  - market-based supply-side policies

Economic inequality and poverty have similar and overlapping causes. The societies that have the most equal distributions of income also tend to have lower levels of poverty (mainly in Nordic countries). The reason is that policies that favour greater income equality overlap with those that reduce poverty. We will examine such policies later in this chapter.

Inequality of opportunity

Opportunity can be defined as a set of circumstances that makes it possible for someone to do something. To understand *inequality of opportunity*, it is useful to compare and contrast it with *economic inequality*.

Economic inequality is concerned with inequalities in standards of living that arise from monetary factors like income and wealth. As such it is concerned with inequalities in *outcomes in standards of living* arising from income and wealth differences.

Inequality of opportunity is concerned with inequalities in *potential outcomes in standards of living* that arise from circumstances that are *beyond one's control.* The World Bank terms it the 'lottery of birth'.[20]

Important circumstances that affect life opportunities and are beyond one's control include such factors as:

- parents' level of education and occupation
- parents' level of income
- place of birth
- gender
- race and ethnicity.

[20] https://blogs.worldbank.org/opendata/data-minute-what-inequality-opportunity

There would be equality of opportunity if everyone began life from a situation where all factors like the above were equal. In such a hypothetical situation, inequalities in *outcomes* would be due to circumstances over which people have control, such as effort in school, effort on the job and hard work.

Yet in the real world, it is apparent that large and growing economic inequalities cannot be explained by differences in circumstances over which people have control, such as effort and hard work. For this reason, economists have recently taken a strong interest in studying the factors that give rise to inequality of opportunity, and the extent to which inequality of opportunity contributes to economic inequality.

An early study of six Latin American countries found that circumstances over which people have no control (father's occupation, parents' education, and region of birth) contributed to economic inequality from at least 25% in Columbia to at least 51% in Guatemala.[21] A larger study of 41 countries found that circumstances contributed 4% in Norway and 40% in Mali.[22]

A more detailed study of the United Kingdom and the United States found that such circumstances as parent's education, time spent with parents, race and childhood behavioural problems are responsible for 31% of inequality in the United Kingdom and 45% in the United States.[23]

Different levels of human capital

Human capital refers to the skills, education and good health that people possess (see Chapter 11). Low levels of education and skills translate into low incomes because there is generally a positive (direct) relationship between skill/educational attainment and income levels. Poor levels of health also lead to low incomes because an unhealthy person is likely to be less productive and therefore more poorly paid. Unskilled people may rely on the minimum wage, which may be insufficient to support a family.

Different levels of resource ownership

Some people inherit, or accumulate through savings from very high incomes, financial capital (cash, stocks and bonds) or other forms of property (such as agricultural land or a home), which gives them both an income advantage as well as increased wealth. Yet many others have no resources to rely on other than their labour, which for numerous reasons (such as low levels of human capital, discrimination and others) may not provide them with an adequate income. People on low incomes often do not own a home and therefore have to pay rent, which may take up a substantial portion of their income without building up their wealth (unlike home-owners gradually paying off a mortgage). Low incomes may mean poor housing, affecting health and further lowering one's income potential, and may even lead to homelessness.

REAL WORLD FOCUS 12.2

Inequality of opportunity in education

On average, children from rich and middle-income families are more likely to do better at school, go to university and have higher incomes as adults. The advantages of children from higher-income families begin from a very young age, as a result of better nutrition and family spending on activities that involve books and a variety of non-school educational activities.

According to a study about the United States, spending per child on education in the poorest fifth of the population increased by about 55% between the mid-1970s and the mid-2000s, while it rose by 155% for the richest fifth over the same period. This, together with additional advantages enjoyed by children of richer families (spending on non-school activities), is increasing the gap in educational attainment between rich and poor.

School systems can often reinforce these trends. For example, schools with children from disadvantaged families often find it difficult to attract qualified teachers, with the result that disadvantaged children receive lower quality education.

21 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4991.2011.00467.x

22 https://openknowledge.worldbank.org/bitstream/handle/10986/12193/wps6304.pdf?sequence=1&isAllowed=y

23 https://www.worldbank.org/en/news/feature/2019/03/28/inequality-of-opportunity-new-measurements-reveal-the-consequences-of-unequal-life-chances

CONTINUED

In all OECD countries, children of parents who did not attend university are themselves less likely to go to university. On average, the proportion of university students of parents who did not themselves attend is about half of what it would have been if all social groups were proportionally represented.

Since people with lower levels of educational attainment have lower levels of income compared to those who are more educated, it follows that inequality of opportunity in education contributes to income inequality. Moreover, the problem is intergenerational, since inequalities are transferred from generation to generation.

Figure 12.6: Ethiopia. Children in school with walls made of clay and straw, with no light or electricity

A study at the London School of Economics has found that the advantages offered by high-income families can persist for more than half a millennium. UK students with names of prominent families could be traced back to the Normans who invaded England in the 11th century, and attended the exclusive universities of Oxford and Cambridge. By contrast, students with 'lower status' surnames attended these universities with far less consistency.

Some researchers also argue that such intergenerational advantages last longer in more unequal societies.

Source: Keeley, B. (2015), Income Inequality: The Gap between Rich and Poor, OECD Insights, OECD Publishing, Paris *https://www.oecd-ilibrary.org/social-issues-migration-health/income-inequality_9789264246010-en*

Applying your skills

Based on your experience you may be able to identify individuals or groups of individuals who face unequal opportunities. Identify how their opportunities differ based on differing circumstances that are beyond their control. Describe some advantages or disadvantages they face due to differing opportunities.

Discrimination

Discrimination is a serious problem both for the individuals involved as well as for the job market. Some social groups (racial and ethnic groups, women) often face discrimination in the job market, with the result that they may receive lower wages than others for the same work, or may find greater difficulty finding work than the worker who does not face discrimination. They often live in poor quality environments and have less access to social services. For example, women are at higher risk of poverty because they are less likely than men to be in paid employment, often do unpaid caring work, they often receive lower pay for the same job, and have lower pensions.

Unequal status and power

Status refers to one's social or professional position in a society. It may be due to level of education, or level of income and wealth, or some form of social arrangement (as in an aristocracy). Status is often closely related to power, as individuals or groups with a high status are also often able to control and influence other people or events. Large inequalities in status can affect economic inequality because people in positions of power may sometimes use this to influence government policies favouring their own interests and hence protecting their incomes and wealth, rather than policies that favour redistribution (to be discussed below. See Real world focus 12.4.).

Government tax and benefits policies

People on low incomes must often rely heavily on transfer payments (see below) and social services and merit goods (health care, education, housing) provided or subsidised by the government, as their incomes are insufficient to purchase these in the market. If these are limited or are reduced by the government, people on low incomes may be forced into poverty by having to purchase these in the market.

In addition, government tax policies play a crucially important role in determining income and wealth distribution. (We will study these later in this chapter.) Tax policies that favour the rich and do not favour redistribution contribute to increasing income and wealth inequalities and poverty. In many countries, particularly developed ones, changing tax policies have contributed to widening income inequalities (see the discussion later in this chapter).

Technological change

While the development of new technologies contributes greatly to improving labour productivity (output per worker; see Chapter 11) and therefore to promoting economic growth, in recent years it has contributed to greater inequality. The reason is that it has eliminated some jobs by replacing human labour by machines (automation). For example, jobs in packaging or manufacturing that require a lot of repetitive work have been replaced by machines that can complete the work faster and more effectively. The result is that wages of low-skill labour whose jobs are being eliminated do not rise much. At the same time, new technologies have created demand for new higher-level skills, meaning that wages of such workers rise faster than those of low-skill workers. As a result, income differences between higher-skill and lower-skill labour are increasing.

Another related factor is that technological change that leads to the replacement of labour by capital (new machines) means that there is an increase in incomes of owners of capital. This results in growing income inequality between workers whose income comes from their labour and the owners of capital who invest in new machines.

Globalisation

Globalisation contributes to the above process. It refers to economic integration on a global scale, involving increasing interconnectedness throughout the world in many areas (trade, finance, investment, people, technology, ideas, knowledge, communications and culture). Increased foreign direct investment (FDI, involving investments by multinational corporations; see Chapter 20) from developed economies increase income inequalities in both developed and developing economies because FDI tends to involve greater demand for skilled rather than unskilled workers, increasing the income differences between the two. In addition, developed economies sometimes offshore certain jobs (relocating them to other countries with lower labour costs), resulting in a lower domestic demand for certain skills.

Market-based supply-side policies

These policies will be studied at length in Chapters 13 and 20, where we will see that in some cases they lead to greater unemployment, or lower incomes for lower-skilled workers, and hence to increased income inequalities and poverty. They have been increasingly used in many countries around the world since the 1980s. Policies such as discouraging trade unions and reduction of the bargaining power of labour, reduction of the minimum wage, and reductions in employment protection have been found to contribute significantly to increasing inequalities.[24]

High abnormal profits of firms with increasing market power (HL only)

Some large firms with market power have been able to earn very high and increasing abnormal profits which transfer income and wealth away from consumers who have to pay higher prices and toward the owners of the firms (see Chapter 7, Section *Evaluating monopoly and comparing with perfect competition*).

Increases in pay of certain occupations

Certain occupations, in particular executives and professionals in the financial sector and non-financial executives have been enjoying huge increases in pay. In the United States, the ratio of pay of CEOs to pay of the average worker increased from 20 to 1 in 1965 to 300 to 1 in 2013.[25]

Unemployment

An unemployed individual receives no income but may receive some unemployment benefits; however; these are generally low relative to income received for work, and

[24] https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf

[25] https://www8.gsb.columbia.edu/faculty/jstiglitz/sites/jstiglitz/files/Inequality%20and%20Economic%20Growth.pdf

in most countries are only provided for limited periods. If unemployment is long term (such as with structural unemployment), then an individual or family is more likely to become poor. The risk of falling into poverty is far greater in single-parent households where the parent is unemployed, or if both heads of a household are unemployed over long periods.

Geography

Some people may live in remote, isolated geographical regions, with limited possibilities for employment, and with limited possibilities to relocate (move) to other more economically active regions (due to poverty, age, or lack of communication and lack of marketable skills); this problem may be especially significant in some rural areas in developing countries.

Age

Older people may receive pensions that are barely enough to cover minimum needs, and in many countries (particularly developing ones) may receive no pension at all if they have been living and working in the informal economy (outside the legally registered economy; see Chapter 19).

Poverty

Poverty itself may become a cause of further poverty. If people do not have access to essential services such as health care, education and housing, a self-perpetuating cycle may be set into motion where low incomes lead to low human capital, and further to low incomes. This is part of the 'poverty cycle', to be studied in Chapter 19.

TEST YOUR UNDERSTANDING 12.3

1. Explain the meaning of inequality of opportunity and using examples discuss how it contributes to economic inequality and poverty.
2. Identify some further causes of economic inequality and poverty, explaining in each case how inequality or poverty are created or worsened.

12.4 The impact of income and wealth inequality

LEARNING OBJECTIVES

After studying this section you will be able to:

- evaluate the impact of income and wealth inequality on (AO3)
  - economic growth
  - standards of living
  - social stability

Economic growth

In Chapter 11 we examined the impact of economic growth on income distribution. Now we look at the impact of income (and wealth) distribution on growth.

There is increasing evidence that high levels of inequality are not good for economic growth. A number of studies confirm this point. For example, a study by the International Monetary Fund (IMF) concludes that lower inequality is linked with faster and more sustained growth, while policies that redistribute income do not generally have negative effects on growth.[26] Another study by the IMF found that periods of growth were more likely to come to an end in countries that have more unequal income distributions.[27]

Yet another study by the IMF based on data from 159 developed and developing countries finds that increases in the share of income of the poor and middle class works to increase growth, whereas an increasing share of the top 20% results in lower growth.[28] According to the study 'if the income share of the top 20% (the rich) increases, then GDP growth actually declines over the medium term' while 'an increase in the income share of the bottom 20% (the poor) is associated with higher GDP growth'.

[26] Redistribution, inequality and growth https://www.imf.org/external/pubs/ft/sdn/2014/sdn1402.pdf

[27] Warning! Inequality may be hazardous to your growth https://blogs.imf.org/2011/04/08/inequality-and-growth

[28] https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2016/12/31/Causes-and-Consequences-of-Income-Inequality-A-Global-Perspective-42986

Reasons why inequality leads to lower growth include the following:

- Greater inequality lowers growth by reducing the ability of lower income households to invest in human and physical capital. For example it leads to lower spending on education, with poor children going to lower-quality schools, which in turn makes it more difficult for them to continue to university. This results in lower labour productivity (output per worker) hence lower growth.
- Countries with higher levels of income inequality have higher levels of inequality of opportunity in education, which is transferred from generation to generation so that children of low-income families are likely to also have low incomes (see also Real world focus 12.2).
- High income inequality may lead to lower growth because the wealthy spend a lower fraction of their incomes than middle income and lower income groups.[29] But the higher savings of higher income groups often leave the country as financial investments abroad, thus reducing resources available for domestic investments.
- The concentration of income and wealth in a few hands results in significant political control and the ability of powerful groups to influence government policies for their own benefit, even though these policies may go against the interests of the whole population. For example, it is considered that the period of higher inequality in developed countries gave rise to activities that led to the global financial crisis of 2008 (such as financial institutions extending too much credit, and reduced government regulation) which greatly reduced rates of growth.[30]
- Significant political control by the rich may also result in less government provision of essential merit goods (education, health care, infrastructure, etc.) which works against the interests of lower income groups and also works against growth. For example, spending on education increases the income-earning potential of the poor, but also leads to greater economic growth by increasing human capital.
- An improved income distribution increases the demand for locally produced goods and services, thus encouraging local production and promoting local employment and investment. With high income inequalities, these potential benefits are lost. This is especially relevant to developing countries.
- Highly unequal income distributions mean that the poor are unable to obtain credit, because they have no collateral as they have no wealth, meaning fewer investments for people on lower incomes, leading to lower growth and development. Also, opportunities to pay for education and health care through borrowing are reduced, leading to lower human capital and lower growth and development.
- A more equal distribution of income leads to greater political stability; highly unequal distributions can lead to social dissatisfaction, unrest and political instability, resulting in lower growth.

Low living standards

This is an obvious consequence arising from low incomes. Low living standards are associated with greater levels of psychological stress, substance abuse, poor nutrition and poor levels of health, all leading to poorer job and income-earning prospects. Low living standards are also a consequence of the factors below:

Lack of access to health care and education. Reduced ability to access health care and education leads to lower human capital formation, lower productivity and lower incomes, possibly resulting in the self-perpetuating poverty cycle noted above (see Chapter 19).

***Higher infant, child and maternal mortality*.** The inability to access needed health care services, as well as poor nutrition for mothers and children lead to large numbers of unnecessary deaths among infants, children and women due to pregnancy-related causes.

Higher levels of preventable diseases. Poor hygiene and nutrition make both children and adults more prone to illnesses.

[29] https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf
[30] https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf

Social problems. These include higher crime rates, drug use, family breakdowns and homelessness.

Inability to realise one's full potential. Due to all of the above people in very low income groups are unable to realise their full potential, leading to a waste of human talent, and in addition to the personal costs, may result in lower economic growth (by adversely affecting the economy's *PPC* or *LRAS* curve).

Social and political stability

High income and wealth inequalities create societies that are polarised and divided, consisting of social groups with different interests that make interactions between them difficult. This leads to a reduced sense of social solidarity and trust in the system, while at the same time the groups at the top begin to feel entitled (that they have rights over others).

The groups at the top begin to have a stronger political influence. The result is that economic inequality leads to political inequality. But those groups at the top with political power influence economic policies (such as tax and social benefits and merit goods policies) in their favour, so that economic inequality becomes even greater. They also influence the political rules of system in order to increase their political power. This results in a vicious circle of greater economic inequality and political inequality.[31]

Growing inequalities increasingly give rise to the feeling that people at or close to the bottom are socially inferior, giving rise to a pronounced sense of dissatisfaction which may eventually pave the way for social instability with possible social conflicts. Governments may further polarise society by serving the interests of their supporters such as lobbyists or big money donors at the expense of the interests of the whole of society. As divisions between social groups widen, it becomes more and more difficult to reach consensus on important challenges.

TEST YOUR UNDERSTANDING 12.4

1. Discuss some of the reasons why high economic inequality is not good for growth.
2. Evaluate the impact of income and wealth inequality on
   a. living standards, and
   b. social stability.

12.5 Policies to reduce income and wealth inequalities and poverty

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between progressive, proportional and regressive taxes (AO2)
- distinguish between (AO2)
  - direct taxes
    - personal income
    - corporate income
    - wealth
  - indirect taxes
- calculate the amount of indirect tax paid from an amount of expenditure, given the indirect tax rate (HL only) (AO4)
- calculate total and average tax rates from data provided (HL only) (AO4)
- explain average and marginal tax rates (HL only) (AO2)
- evaluate the role of different types of taxes in reducing income and wealth inequalities and poverty (AO3)
- explain and evaluate further policies to reduce income and wealth inequalities and poverty including (AO3)
  - policies to reduce inequalities of opportunity such as investment in human capital
  - transfer payments
  - targeted spending on goods and services
  - universal basic income
  - policies to reduce discrimination
  - minimum wages

[31] https://www8.gsb.columbia.edu/faculty/jstiglitz/sites/jstiglitz/files/Income%20Inequality%20PPT.pdf

Taxation policies to reduce income and wealth inequalities and poverty

Income and wealth inequalities and poverty as we have seen are major global concerns. To make matters worse income and wealth inequalities are increasing in many countries around the world. In this section we will consider a number of policy tools governments can use to promote equity (or a greater degree of income or wealth equality) and alleviation of poverty.

The role of taxation in promoting (or worsening) equity (income redistribution)

One of the most important instruments for income and wealth redistribution is taxation, because it can lower inequalities by taking more taxes from the rich than from the poor.

Taxes are the most important source of government revenues, and provide the funds for many purposes, such as public goods, transfer payments and merit goods, correcting externalities, providing subsidies, changing the allocation of resources, changing the distribution of income, and many more.

There are two very broad categories of taxes, direct taxes and indirect taxes.

Direct taxes

Direct taxes are taxes paid directly to the government tax authorities by the taxpayer. The most important kinds of direct taxes include the following:

- **Personal income taxes.** This is the most important source of government tax revenues in many countries (especially developed countries), and involves taxes paid by households or individuals in households. They are paid on all forms of income, including wages, rental income, interest income and dividends (which are income from ownership of shares in a company, and are therefore income from profits).
- **Corporate income taxes.** Corporations are businesses (firms) that have formed a legal body called a 'corporation' that is legally separate from its owners. Corporate income taxes are taxes on the profits of corporations.
- **Wealth taxes.** These are taxes on the ownership of assets. Two common wealth taxes are *property taxes*, based on the value of property owned, and *inheritance taxes*, based on the value of property inherited.

The revenue collected from all of the above forms of taxation is paid into the government's budget and is used to finance a broad variety of government expenditures. In contrast to these, there is an additional form of direct taxation:

- **Social insurance (social security) contributions or payroll taxes.** These taxes are paid by workers and their employers (who pay on behalf of their employees). The revenues from these taxes are not paid into the government's budget, but rather into specific funds, and are used to finance specific expenditures, such as pensions, social insurance and health care (in some countries).

Indirect taxes

Indirect taxes are taxes on spending on goods and services, discussed in Chapters 4 and 5. They are called indirect because, while consumers are the ultimate payers of a part of these taxes, they pay indirectly through the suppliers of the good or service purchased (the suppliers may be the producers, the retailers, or generally the sellers). The most important kinds of indirect taxes include the following:

- **General expenditure taxes, also known as sales taxes**. These are taxes on spending or sales of goods and services. In the United States (where they are known as 'sales taxes'), they are a fixed percentage of the retail price of goods and services. The corresponding tax used in the European Union and many other countries around the world is the 'value added tax' (VAT). VAT differs from a sales tax in that it is a tax paid on the value added by each producer in the production process.[32] In practice, many countries that use the VAT as well as states in the United States exempt or exclude certain goods and services from payment of taxes on the grounds of equity (for example, goods

[32] 'Value added' refers to the value of a firm's output minus the value of its inputs. If there is more than one stage in the production process, the firm pays VAT for each stage in which value is added. Thus the total VAT paid by the firm is the sum of the value added taxes paid at each stage in the production process. When the good or service reaches the marketplace, its price includes the VAT that has been paid by all the firms involved in its production. Each country in the European Union has its own particular VAT rates.

	Proportional taxation		Progressive taxation				Regressive taxation	
Income €	Tax rate (proportional)	Amount of tax €	Tax rate (mildly progressive)	Amount of tax €	Tax rate (strongly progressive)	Amount of tax €	Tax rate	Amount of tax €
10 000	15%	1500	15%	1500	15%	1500	15%	1500
20 000	15%	3000	20%	4000	25%	5000	10%	2000
30 000	15%	4500	25%	7500	40%	12 000	5%	1500

Table 12.7: Progressive, proportional and regressive taxation

and services like food, pharmaceuticals, rents on housing, and others); in some cases there may be different rates for different types of goods and services on the grounds of equity.

- **Excise taxes.** In contrast to expenditure/sales taxes, excise taxes are taxes paid on specific goods and services, such as cigarettes and petrol (gasoline). (See Chapters 4 and 5.)
- **Customs duties, also known as tariffs.** Customs duties or tariffs are a type of tax applied on imports of foreign goods into a country. There are two main reasons why governments levy tariffs. One is to keep imports out of the country by making them more expensive to consumers, and the other is to raise tax revenues. We will study tariffs in Chapter 14.

Tax systems vary enormously from country to country around the world, and there are no two countries that have the same tax system. While most countries have a mix of direct and indirect taxes, they vary with regard to the degree of reliance on each of these, as well as on the particular mix of types of taxes within each group. In addition, they can vary enormously with respect to tax rates and many other technical details.

Understanding the principles of proportional, progressive and regressive taxation

Taxes can be defined as being proportional, progressive or regressive according to the relationship between income and the fraction of income paid as tax. The fraction of income paid as tax, in percentage terms, is referred to as a *tax rate*.

- **Proportional taxation:** as income increases, the fraction of income paid as taxes remains constant; there is a constant tax rate.
- **Progressive taxation:** as income increases, the fraction of income paid as taxes increases; there is an increasing tax rate.
- **Regressive taxation:** as income increases, the fraction of income paid as taxes decreases; there is a decreasing tax rate.

These relationships are illustrated in the numerical examples appearing in Table 12.7. The table shows three possible income levels, and four hypothetical taxation systems with differing tax rates. In each case, the amount of tax is calculated by multiplying the amount of income times the corresponding tax rate. In the case of proportional taxation, the tax rate remains constant (at 15%) for all income levels; therefore, the amount of tax paid increases in the same proportion as income. In the case of progressive taxation, two examples are presented. In both, the tax rate increases as income increases; as a result, the amount of tax paid increases more than in proportion to income. The first example is 'mildly progressive', meaning that the tax rate increases slowly as income increases. In the 'strongly progressive' example, the tax rate increases much more rapidly.

Ignoring regressive taxation for the moment, let's compare the amount of tax paid in the cases of proportional, mildly progressive and strongly progressive taxation. In all three cases, as income increases, the amount of tax paid becomes larger. Which of the three is the most 'fair'? (All three satisfy the principle of vertical equity, according to which those with greater income should pay a larger tax than those with less income; see Theory of knowledge feature 12.2.) This is impossible to answer using economic reasoning, because fairness is a normative question that depends on value judgments.

The only thing that can be said with certainty is that the more progressive a tax system, the more equal (or less unequal) the after-tax distribution of income becomes. We can see why by examining the figures in the table. In all three taxation systems, the taxpayer faces the same tax rate and the same amount of tax for income of €10 000. However, as income increases to €20 000 and then to €30 000, the amount of tax paid increases more rapidly in the mildly progressive system and even more rapidly in the strongly progressive system. By shrinking the difference between the high and low income levels, progressive taxation achieves a more equal income distribution than proportional taxation; and the more progressive the tax system, the greater the income equality achieved.

In regressive taxation, the tax rate decreases as income increases; the proportion of income paid in tax falls as income rises. Regressive taxation therefore makes income distribution less equal.

Why indirect taxes are regressive

Although progressive, proportional and regressive taxation are defined relative to income, they do not apply just to income taxes, but to all types of taxes, whether direct or indirect. This is because taxes are paid out of income, and can therefore be compared with income. To see whether a tax is progressive, proportional or regressive, it can be calculated as a fraction of the income out of which it is paid. For example, corporate income taxes are usually proportional. Social insurance contributions are also usually proportional (for both workers and their employers). Indirect taxes (of all types) are regressive.

To see why indirect taxes are regressive, consider two individuals, XX and XY. Each one buys a car for $10 000 (without tax), on which there is an indirect tax (VAT, or sales tax, or tariff) of 10%, amounting to $1000. Individual XY has an annual income of $10 000 and individual XX has an annual income of $20 000. The $1000 of tax is 10% of individual XY's income, and it is 5% of individual XX's income. In other words, as income increases from $10 000 to $20 000, the fraction of income paid on the indirect tax decreases from 10% to 5%. This is the definition of regressive taxation. Since all indirect taxes work on the same principle, we can conclude that indirect taxes are regressive. *It follows that indirect taxes are inconsistent with the objective of a more equal distribution of income*.

In the real world, some goods and services considered to be necessities, such as food and medicines, may be exempted from general expenditure or sales taxes. This makes indirect taxes somewhat less regressive, though they remain regressive overall.

The tax systems of virtually all countries consist of many different types of taxes. Whether the overall tax system of a country is more progressive or regressive depends on the mix of taxes and tax rates, and their relative importance as sources of government revenue. For example, a country that relies relatively more on progressive income taxes is likely to have a tax system that is relatively more progressive compared to a country that relies relatively more on indirect taxes.

The more progressive a tax system, the more equal is the after-tax distribution of income compared to the pre-tax distribution of income. Regressive tax systems tend to make the distribution of income less equal.

TEST YOUR UNDERSTANDING 12.5

1 Distinguish between direct and indirect taxes, providing examples of each.

2 Define and explain, using examples, the difference between proportional, progressive and regressive taxation.

3 Explain how income taxes can be used to lower income inequalities.

4 a Explain why indirect taxes are regressive.

b Outline the reason for excluding some goods and services considered to be necessities from payment of indirect taxes (such as VAT and sales taxes).

c Describe the impact this has on how regressive the indirect taxes are.

5 Outline why the issue of using taxes to redistribute income is partly a normative issue.

THEORY OF KNOWLEDGE 12.2

Principles of equity for taxation

Equity refers to the idea of being fair or just. The design of a tax system requires some principles of equity if it is to be considered fair or just. There are two main principles of equity in taxation, explained below.

The benefits-received principle

According to this principle, consumers and firms should pay taxes on goods and services provided by the government in accordance with the use they make of them. The idea here is that such goods and services should be paid for by their users in the same way that goods and services provided by the market are paid for: the tax serves a similar function as price in the market, and tax revenues are used to pay (at least partly) for provision of the goods and services by the government. Examples of goods paid for (taxed) this way include parks that have admission fees, roads and bridges that charge tolls, roads and highways financed by petrol (gasoline) taxes, and publicly financed schools paid for by property taxes.

In each case, the benefits principle is used, even though it is not always the case that the payer of the tax is the one who immediately benefits from the use of the good or service. For example, property taxes used to pay for publicly-financed schools are based on the assumption that people who live in a particular area will send their children to the local school; the petrol (gasoline) tax used to pay for highway construction and maintenance presupposes that those who buy petrol (gasoline) are also those who will benefit from the use of the highway.

The benefits principle is considered 'fair' in the sense that people pay for benefits they receive, and do not pay for benefits that are enjoyed by others. However, it cannot be used as a general principle for taxation for two main reasons. One is that it is very difficult, if at all possible, to identify who receives the benefits of numerous public goods like national defence, police protection and the court system. The other is that it does not allow for any redistribution, such as that provided by transfer payments.

The ability-to-pay principle

According to this principle, taxes should be levied on individuals or households according to their ability to pay a tax, or in accordance with their income or wealth. The 'ability-to-pay' principle is derived from the ideas on distributive justice (fairness) of the famous Greek philosopher Aristotle, who distinguished between horizontal and vertical equity. According to the idea of horizontal equity, people who are equal with respect to certain characteristics should be treated equally. When applied to a tax system, this means that people with similar income or wealth (or a similar ability to pay) should be taxed by the same amounts. According to the idea of vertical equity, people who are unequal with respect to certain characteristics should be treated unequally, and in proportion to their inequalities. In a tax system, since many people are

CONTINUED

unequal with respect to income and wealth (ability to pay), those with greater income and wealth should pay a larger tax than those with less income and wealth. If a tax system is designed with these ideas in mind, the distribution of income will become more equal (or less unequal).

The ability-to-pay principle is considered 'fair' because people with higher incomes pay more tax than people with lower incomes. It also helps solve the problem of redistribution through such methods as transfer payments. Whereas it is widely accepted, and is used in most countries as the basis for designing income tax systems, in practice it is very difficult to reach agreement about how the concepts of vertical and horizontal equity should be interpreted. Even if it is agreed that the rich should pay more than the poor, the question is, how much more? This is a highly controversial question, and the answer depends very much on political beliefs and value systems (normative issues).

Thinking points

- Do you think it is important for a tax system to be based on a principle of equity that is generally accepted by the members of a society?
- Examine the benefits-received and the ability-to-pay principles from the perspective of equity. Which do you think is more 'fair'? How would you justify your argument?
- Do the people of a society have a moral obligation to pay taxes? What if they do not agree with the equity principle of the tax system?

Achieving progressivity through proportional taxation (Supplementary material)

If you are interested you can read about this in the digital coursebook as Supplementary material.

Calculations of direct and indirect taxes (HL only)

How to calculate total income taxes and average tax rates

To see how income taxes are calculated in the real world, we will make a distinction between marginal and average tax rates. A **marginal tax rate** is defined as the tax rate paid on additional income, or on the last amount of tax paid, expressed as a percentage. An **average tax rate** is tax paid divided by total income, also expressed as a percentage. All the tax rates appearing in Table 12.7 are average tax rates.

In the real world, income taxes in a progressive tax system are calculated using successive layers of income and applying a different tax rate to each layer. The layers of income are called tax brackets, and the corresponding tax rates are the marginal tax rates. A numerical example will help you understand this. Column 1 of Table 12.8 shows tax brackets (the layers of income) in a hypothetical economy, and column 2 gives us the marginal tax rate that applies to each bracket.

Suppose we want to calculate the amount of income tax paid on an annual income of \$59 000. The total tax paid will be 0 for the first \$10 000 of this income; 9% on income between \$10 001 and \$25 000; 22% on income between \$25 001 and \$55 000; and finally 40% on income between \$55 001 and \$59 000. We calculate the total tax paid as follows.

$(0 \times \$10\,000) + (0.09 \times \$15\,000) + (0.22 \times \$30\,000) + (0.40 \times \$4000) = 0 + \$1350 + \$6600 + \$1600 = \$9550$

What is the average tax rate for the income of \$59 000? It is total tax paid divided by income expressed as a percentage, or

$$\frac{\$9550}{\$59\,000} = 0.162, \text{ or } 16.2\%.$$

Using the information in Table 12.8, let's now calculate the income tax paid on income of \$175 000.

$(0 \times \$10\,000) + (0.09 \times \$15\,000) + (0.22 \times \$30\,000 + (0.40 \times \$60\,000) + (0.55 \times \$60\,000) = 0 + \$1350 + \$6600 + \$24\,000 + \$33\,000 = \$64\,950 =$ income tax paid

The average tax rate on income of $175 000 is 37.1%

i.e. $\frac{\$64\,950}{\$175\,000} = 0.371, \text{or } 37.1\%$.

Comparing the average tax rate for income of $59 000 (16.2%) with the average tax rate for income of $175 000 (37.1%), we see that the higher income has a higher average tax rate, just as expected since this is a progressive tax system.

1 Annual income ($)	2 Marginal tax rate (%)
0–10 000	0
10 001–25 000	9
25 001–55 000	22
55 001–115 000	40
115 001 or more	55

Table 12.8: Income taxes with increasing marginal tax rates

Calculations involving indirect taxes

Suppose a family with an annual income of €60 000 pays income taxes of €15 000. Its *disposable income* is €45 000 (disposable income is after tax income so it is 60 000 − 15 000). Its *average rate of income tax* is

$$\frac{15\,000}{60\,000} = 25.0\%$$

Suppose the family spends all of its disposable income. This spending includes an indirect tax that is imposed on purchases of all goods and services of 12.5%. Therefore the actual amount spent on the purchase will be less than €45 000. The question we want to answer is *how much indirect tax* this family pays.

You might be tempted to say that the amount paid is 12.5% of €45 000 which is €5625 (= 0.125 × 45 000). However this would be incorrect.[33]

To find the indirect tax paid, let Z = the amount of spending on goods and services themselves not including the indirect tax.

It follows then that:

Z + 0.125 Z = 45 000 or more simply 1.125 Z = 45 000. Solving for Z we have

$$Z = \frac{45\,000}{1.125} = 40\,000.$$

There are two ways we can now find the amount of tax. One way is to subtract spending before the tax from spending after the tax, so that:

€45 000 − €40 000 = €5000 = amount of spending on the indirect tax.

An alternative way is to take 12.5% of 40 000 or 0.125 × 40 000 = €5000.

More generally we can say that since the amount of spending on the indirect tax = 0.125 Z, and since

$Z = \frac{45\,000}{1.125}$ it follows that

amount of spending on the indirect tax =

$$0.125 \times \frac{45\,000}{1.125} = €5000$$

> Generalising, we can say that if *S* = the amount available for purchases and *r* = the rate of indirect tax, the amount of spending on the
>
> $$\text{indirect tax} = \frac{r}{100} \times \frac{S}{1 + \frac{r}{100}}$$
>
> To convince yourself, you can use the example above by letting S = €45000 and r = 12.5%. You will find exactly the same result as above.

It is now possible to calculate the *average rate of indirect tax* for this family.

$$\text{average rate of indirect tax} = \frac{5000}{60\,000} = 0.0833 \text{ or } 8.33\%.$$

Suppose we now want to find the total average rate of tax of this family, that includes both the direct and the indirect tax. To do this we add the amount of direct tax to the amount of indirect tax and divide by total income

$$\text{total average tax rate} = \frac{15\,000 + 5000}{60\,000} = \frac{20000}{60000}$$
$$= 0.3333 \text{ or } 33.33\%$$

Note that we can also find this by adding together the two average tax rates that we found above: average rate of income tax + average rate of indirect tax = 25.0% + 8.33 = 33.33%

[33] To see why this is so note that if €5625 is paid as indirect tax, this means that €39 375 (= 45 000 − 5625) is the amount of spending to buy goods and services. Since there is a 12.5% tax on spending to buy all goods and services, this means that the family must pay 0.125 × €39 375 = €4921.87 on tax, making total spending = €39 375 + €4921.87 = €44 296.87 which is not possible since total spending including tax is €45 000.

TEST YOUR UNDERSTANDING 12.6

The following questions are based on the data in Table 12.8.

1 Calculate the amount of income tax paid by families with annual income levels of

   a $6500,

   b $15 700,

   c $31 000,

   d $48 000, and

   e $120 000.

2 For each of the items in question 1,

   a calculate each family's average income tax rate,

   b outline what happens to the average tax rate as income increases and explain why, and

   c state each family's marginal tax rate.

3 Suppose that the family with income of $48 000 spends 85% of its disposable (after tax) income (it saves 15% of it). Suppose, too, that the family pays a 17% value added tax (VAT) on its spending on all good and services.

   a Calculate the amount of indirect tax paid.

   b Calculate the amount of indirect tax paid as a percentage of income (the average indirect tax rate).

   c Using your answer to question 2, calculate the total average tax rate including both direct and indirect tax.

This means that this family is paying one-third of its total income, or 33.33% as taxes, including both direct and indirect.

Evaluating taxes as a policy for redistribution

In advanced economies, an estimated 25% of redistribution occurs through the tax system (the remaining 75% occurring through benefits like transfer payments; see below).

A long-standing debate among economists has concerned the possible conflict between economic growth and equity, interpreted to mean greater economic equality. In Section 12.4 we saw that there is strong evidence that a high level of economic inequality is not good for growth. Therefore it would seem to follow that tax systems that are designed to reduce income inequalities would be good for growth. Such tax systems would favour direct taxes of all types that are progressive (personal income taxes, corporation taxes and wealth taxes), since the more progressive a tax system, the greater the fraction of income, profit or wealth that is taxed away, and therefore the greater the resulting equality in distribution. This line of reasoning would also suggest there should be less reliance on indirect taxes, as these are regressive.

Yet many economists would disagree on the grounds that progressive taxes have disincentive effects. It is argued that *income taxes* reduce after-tax income, acting as a disincentive to work, as well as to save, particularly among high income earners who would be taxed more heavily. They have the effect of reducing the quantity of labour offered in the market and also reduce savings. Lower savings has a negative effect on investment, and therefore on production of new capital goods. In addition, *corporation taxes*, which are taxes on the profits of corporations, have the effect of reducing the incentives to invest, also resulting in lower production of new capital goods. Lower quantities of labour and capital in turn translate into lower rates of growth for the economy. Some economists further argue that *wealth taxes* also have negative effects on efficiency and innovation by reducing incentives to invest in productive capital, thus also leading to lower growth.

This thinking is very much in line with market-based supply-side policies that we will study in Chapter 13, which has become increasingly popular in many countries since the 1980s. This may explain the trends in reductions of certain kinds of direct taxes in many countries in the last decades. Since the 1980s the role of taxes has weakened as a method for redistribution. This has been due to declining tax rates in many countries, which have contributed to increasing economic inequality.

For example, of the 36 countries of the OECD, twelve countries had wealth taxes in 1990, and by 2017 these had been reduced to only four countries.[34] In the case of inheritance taxes, the proportion of government revenues raised by this type of tax have fallen by three fifths since the 1960s in OECD countries. Corporation taxes have also been falling in many countries. In 1990 the average corporation tax in the G20 countries[35] was 40%. By 2015 it had fallen to 28.7%. In addition, multinational corporations make increasing use of tax havens (countries where corporations do not have to pay taxes) as well as other methods to avoid paying taxes. The IMF estimates that tax revenues from multinational corporations worth as much as 1% of GDP of OECD countries are lost each year.[36]

Yet there is no evidence that lower taxes have the effects that their proponents claim. Lower income taxes need not lead to more work, they may simply lead to more leisure time (time away from work). Lower corporation taxes need not lead to greater investment, in fact a cut in the corporation tax that took place in the United States in 2017 led to *less* investment by corporations (see Real world focus 12.4 and especially 13.2 in Chapter 13). Moreover, contrary to the above claims, certain wealth taxes, including taxes on real estate and land are both equitable and efficient.[37]

Instead, the above trends have led to increasing economic inequality.[38]

In view of the above, there are many prominent economists who argue that the best way to reverse the trend toward increased wealth inequality is by taxing corporations and wealth, including imposing inheritance taxes. For example, a major study by the OECD[39] argues that 'there is a strong case for addressing wealth inequality through the tax system.' It recommends broad-based capital income taxes (income from financial investments) which should be complemented by a form of wealth taxes, preferably inheritance taxes.

Moreover, the International Monetary Fund[40] (IMF, see Chapter 20) argues that a number of different wealth taxes should be considered to address the problem of increasing inequality. Further, according to the IMF declining progressivity in income taxes of many countries in recent decades should be reversed, without fear that this will reduce growth, because there is no evidence that higher income taxes will have negative effects on growth.

Other policies to reduce income and wealth inequalities and poverty

In addition to taxes there are a number of further redistribution policies that governments can use to address economic inequality and poverty.

Policies to reduce inequality of opportunity: investment in human capital

Inequality of opportunity is a major factor leading to income and wealth inequality. A key policy to tackle this involves investment in human capital including access to high quality education and health care.

In education, disadvantaged groups have much lower educational attainment. For example, in sub-Saharan Africa, where the differences between high and low income people are the most pronounced, almost 60% of the poorest quintile (20%) of the youth population aged 10–24 have less than four years of schooling, compared to 15% of the richest quintile.[41]

Regarding access to health care, there are significant differences between the richest and poorest quintiles everywhere in the world, though they are more pronounced in developing countries. For example, infant mortality rates in developing countries on average are about 67 per 1000 live births for the lowest quintile compared to 48 for the highest quintile. In emerging market economies the corresponding figures are 34% and 18%.[42]

People with deprivations in areas such as education and health are unable to offer their children opportunities available to higher income families. Therefore investment by the government in human capital to ensure universal access to education and health care is imperative.

[34] https://www.imf.org/en/Publications/FM/Issues/2017/10/05/fiscal-monitor-october-2017

[35] An international forum consisting of nineteen countries and the European Union intended to promote international cooperation.

[36] https://www-cdn.oxfam.org/s3fs-public/file_attachments/bp-economy-for-99-percent-160117-en.pdf

[37] https://www.imf.org/en/Publications/FM/Issues/2017/10/05/fiscal-monitor-october-2017

[38] https://www.imf.org/en/Publications/FM/Issues/2017/10/05/fiscal-monitor-october-2017

[39] https://www.oecd.org/tax/tax-policy/role-and-design-of-net-wealth-taxes-in-the-OECD-summary.pdf

[40] https://www.imf.org/en/Publications/FM/Issues/2017/10/05/fiscal-monitor-october-2017

[41] https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf

[42] https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf

To reach people on very low incomes these services should be provided free of charge rather than rely on out-of-pocket (private) payments or private provision, as such services are beyond the reach of the poor.

Transfer payments

Transfer payments (*cash transfers*) are payments made by the government to individuals specifically for the purpose of redistributing income away from certain groups and towards other groups; they transfer income from those who work and pay taxes towards those who need assistance. The groups of people who receive the transfer payments may include older people, sick people, very poor people, children of poor families, unemployed people and others; in their entirety they are referred to as *vulnerable groups*. Transfer payments include old age pensions, disability pensions, unemployment benefits, war veterans' benefits, maternity benefits, child allowances, housing benefits for the poor, student grants, and many more. Transfer payments are made possible by taxes collected by the government. A portion of taxes paid to the government by the working population is used to make transfer payments to vulnerable groups, thereby achieving some income redistribution.

Transfer payments are used in many countries around the world, both developed and developing. As many as 60 or more developing countries use *conditional cash transfers*, meaning that these are granted to poor households on condition that they meet certain requirements, usually linked with children's education and health care. They have become a major anti-poverty tool and have also been introduced in the United States. They are especially important as they focus on the extremely important objective of building human capital

Transfer payments play a major role in improving income distribution. In developed countries they contribute an estimated 75% to improved distribution (the 25% being due to the tax system). A disadvantage of transfer payments involves the burden on the government budget and the opportunity costs (sacrificed alternative government spending). Also, some argue that transfer payments create incentives for people not to work, however there are serious questions about the extent to which this argument is valid (see Real world focus 12.3).

Targeted government spending on goods and services

Governments spend to provide *merit goods*, which are goods that are beneficial for consumers, often with positive consumption externalities, that are underprovided by the market and underconsumed (see Chapter 6). In the absence of government intervention, two of the most important merit goods that would be underconsumed due to low incomes and poverty would be education and health care. Education and health care are so important that they are often viewed as fundamental human rights.

This means that it is not enough for governments just to provide education and health care (to supplement the insufficient quantities provided by the market). Governments must also ensure that these are affordable for very low income groups. This can be accomplished when governments offer education and health care services that are free (or nearly free) of charge to consumers. Governments may also provide subsidies to private providers to increase supply (see Chapter 6). In addition, education and health care can be made more affordable through transfer payments.

Other merit goods that are especially important in developing countries where there are large groups of people on very low incomes include *infrastructure*, which consists of numerous kinds of physical capital, such as clean water supplies, sanitation and sewerage. Offering these at zero or low prices makes them affordable to poor people who would otherwise be unable to pay for them. (Infrastructure will be discussed in more detail in Chapter 19.)

Whatever the merit good, the government uses tax revenues to provide the good in larger quantities than the market would have provided, and additionally to make it available at very low (or zero) prices, thus offering substantial amounts of redistribution. In addition, as we know from Chapter 6, provision of merit goods involves correction of positive consumption externalities, which are a type of market failure, and attempts to correct the market's underallocation of resources.

On the negative side, merit good provision is a burden on the government budget and entails opportunity costs in terms of foregone alternatives.

Universal basic income

Universal basic income is a method intended to provide residents in a country with a sum of money that they would receive regardless of any other income they may have. Its purpose is to reduce income inequalities and poverty. It is based on the principle that everyone in a society is entitled to a basic income, regardless whether or not they work for it or under what circumstances they have been born into. The idea has become more popular in view of recognition of growing income inequalities, as well as fears that new technologies will increasingly give rise to job losses leading to greater poverty.

Financing would come partly from a tax and possibly too from savings from cutting other social service programmes.

Supporters of this idea claim that it will be effective in reducing poverty, and it would be administratively very simple to carry out. It would support diverse groups like students in university, young couples looking to start families, unpaid care workers, and budding entrepreneurs with no other source of income. It would provide a better balance between workers and employers by giving workers more freedom to leave jobs they are unhappy with. It would also stabilise the economy during a recession by providing income to the unemployed (like an automatic stabiliser; see Chapter 13).

Opponents claim that it would be too expensive, since all households including middle-income and high-income ones that do not need it would also receive it. It has been argued that it could be financed only if there were serious cuts in other social services, such as transfer payments to vulnerable groups (see above). Yet it may be unreasonable to cut money going to people who need it in order to provide an income to everyone including those who do not need it. In view of this, an alternative that has been proposed would be to offer a basic income only to people whose income falls below a certain amount.

In addition there are worries that some people may lose the incentive to work, though trials carried out in Finland, North Carolina (United States) and Seattle do not provide evidence that this occurs. There are also opportunity costs of government spending due to diversion of funds from other priority areas like health care and education.

An interesting case which appears to be successful is the 'Alaska Permanent Fund' where every citizen of the state receives money every year (though it is not called a universal basic income). While many experiments have been done there is no country as yet at the time of writing where a universal basic income has been adopted on a national level.

Policies to reduce discrimination

Countries around the world usually have legislation that forbids discrimination in the workplace. This is essential to ensure that discrimination does not occur. Governments must further ensure that employers are informed about laws on discrimination. In addition to laws, efforts must be made to educate employers and workers on the benefits of multiculturalism. Additional measures include management training on anti-discrimination practices, and communication of anti-discrimination policy to employees.

However, employers may not always comply with the law. In some cases they may not be sufficiently concerned about discrimination to undertake the variety of measures above to prevent it.

Government intervention in markets: minimum wages and price controls

Price controls affect the distribution of income. One type of price control involves:

- *minimum wage legislation*, which by setting a legal minimum wage, raises the lowest permissible wage above the equilibrium market level, thereby raising the wages of low-income (and usually unskilled) workers. Whereas standard economic theory predicts this will create unemployment, actual practice indicates that this is most often not the case, and therefore works to improve income distribution while also usually increasing employment (see Chapter 4).

In addition, governments may use price controls including:

- food price ceilings that set maximum prices for certain food products (prices below the market-determined equilibrium price), making food more affordable for low-income groups, or rent controls that set maximum rents to support low income people
- price floors for farmers that set legal minimum prices for certain agricultural products (often involving government purchases of the resulting surpluses), raising their prices above the equilibrium market price in order to support farmers' incomes.

Both these types of price controls lead to allocative inefficiency and a loss of social surplus (see Chapter 4).

TEST YOUR UNDERSTANDING 12.7

1 Discuss the importance of investment in human capital as a method to overcome inequalities of opportunity.

2 Using examples, explain how transfer payments and targeted spending on goods and services can help redistribute income.

3 Explain how universal basic income is intended to work.

4 Explain what is likely to happen to a country's Lorenz curve and Gini coefficient if its government:

   a increases income tax rates applied to higher incomes

   b places a greater emphasis on indirect taxes relative to direct taxes as sources of tax revenues

   c reduces transfer payments

   d introduces a system of free education and health care

   e enacts legislation to reduce discrimination.

Discuss the advantages and disadvantages of

   a taxes,

   b transfer payments,

   c targeted spending on goods and services,

   d universal basic income,

   e discrimination policies, and

   f government intervention in the form of price controls including minimum wages.

REAL WORLD FOCUS 12.3

The role of taxes and benefits in UK income redistribution

In the United Kingdom, the top 20% (quintile) of the population has an income that is twelve times the income of the bottom quintile before taxes and social benefits. After taxes and benefits this number falls to five times.

While progressive taxes play a role, benefits are responsible for the largest part in the drop in inequality.

The bottom 20% (quintile) of the population receives sixteen times more in benefits as a share of income than the top quintile does.

Figure 12.7: Jaywick, England. The most deprived area in England

But the top quintile pays only 2.7 times as much direct tax as a share of income as the bottom quintile. Note that this does not account for the effects of indirect taxes, which are regressive.

In 2017, the government imposed a limit on benefits, which restricted these to the first two children only. Within the first two years of the cuts, an estimated 600 000 children were affected. It is expected that an additional 300 000 children will be pushed into poverty by 2024. The policy has been justified on the grounds that the cuts will provide an incentive for people on benefits to work. According to official figures however most of the people affected are already working.[43]

[43] https://www.theguardian.com/society/2019/jun/26/two-child-benefit-limit-pushes-families-further-into-poverty-study

CONTINUED

In 2019, there were proposals for a tax cut that would benefit higher income earners, as well as proposals to reduce the corporation tax even though this is already one of the lowest among developed countries.

Source: *www.theguardian.com/*

Applying your skills

1. Draw two Lorenz curves showing the United Kingdom's income distribution before and after taxes and benefits.
2. Explain why benefits make a greater contribution to income redistribution than taxes in the United Kingdom.
3. Discuss the likely effects on the United Kingdom's income and wealth distribution of the cuts in child benefits and proposed tax changes.

REAL WORLD FOCUS 12.4

Increasing market concentration, inequality and low growth (HL only)

Joseph Stiglitz, a Nobel Prize winning economist, points out that inequality in the United States is at its highest since 1928, GDP growth has been very low in comparison with the decades after the Second World War, and market concentration is growing (see Chapter 7, Real world focus 7.2). Stiglitz sees these three trends as being interconnected.

Market power has increased because business leaders make great efforts to create barriers to competition; as one put it 'competition is for losers'. Competition laws have not been updated and are not being strictly enforced. With growing market power large firms 'exploit their customers' through higher prices and 'squeeze their employees, whose own bargaining power and legal protections are being weakened'. (Stiglitz is referring here to market-based supply-side policies; see the section above under *Causes of economic inequality and poverty*.) Corporate profits are therefore increasing. At the same time that workers' pay is stagnant, CEOs and senior executives are increasing their own pay, without using higher profits to increase investment.

Recent tax cuts of 2017 in the United States, which included cuts in the corporation tax, are expected to favour the rich while hurting the poor.[44] To make matters worse, Stiglitz notes that corporate executives used most of the benefits of the tax cuts to enrich themselves rather than to invest more.

Stiglitz argues that increasing market power of large firms increases their ability to influence America's policies. 'And as the system has become more rigged in business's favour, it has become much harder for ordinary citizens to seek redress for mistreatment or abuse.'

Stiglitz is concerned about the effects of rising inequality which leads to falling aggregate demand as the rich consume a smaller fraction of income. At the same time, market power reduces the incentive to invest and innovate. 'Political investments in getting lower taxes yield far higher returns than real investments in plant and equipment.' The table below shows income growth for different income groups over a 35-year period.

[44] https://www.vox.com/policy-and-politics/2018/2/26/17044440/gop-tax-cut-tax-reform-law-chart-regressive-rich-poor-graph

CONTINUED

Share of population	Percentage increase
Top 0.1%	236%
Top 1%	142%
Top 20%	95%
Fourth 20%	28%
Third 20%	28%
Second 20%	28%
Bottom 20%	26%

Table 12.9: US income growth in real terms, 1979–2014[45]

Source: *www8.gsb.columbia.edu/*

Figure 12.8: Social inequality between rich and ordinary people

Applying your skills

1 Identify possible 'barriers to entry' studied in Chapter 7 and explain how they may be linked to rising income and wealth inequalities.

2 Explain what Stiglitz means by 'the system has become more rigged', and how it relates to rising income and wealth inequalities.

3 Use an *AD-AS* diagram to help explain 'the effects of rising inequality' due to the difference in spending patterns of rich versus low- and middle-income households.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter

1 Research the definition of 'poverty' in the country you live in or a country of your choice. What income is required to be above poverty line? Discuss with your classmates what relative poverty looks like where you live. What kinds of goods and services are considered to be things that most of the population enjoys?

2 Investigate policies to reduce poverty and inequalities in a country of your choice. Consider the extent to which these have been successful, and possible ways that they might be improved.

3 Research three countries where universal basic income experiments have been carried out. Evaluate the effectiveness of each given their respective results.

4 Identify a country of your choice and research how its income and wealth distributions have changed over the last couple of decades. What are the factors that have contributed to the changes you observe?

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

[45] https://www.cbo.gov/system/files/115th-congress-2017-2018/reports/53597-distribution-household-income-2014.pdf
https://www8.gsb.columbia.edu/faculty/jstiglitz/sites/jstiglitz/files/Inequality%20and%20Economic%20Growth.pdf

Real world issue 2: What can governments do to manage their economy and how effective are their policies?

CONCEPTUAL UNDERSTANDINGS

1. Macroeconomic objectives are pursued by government intervention through a *choice* of policies.
2. The effectiveness of government policies is influenced by political, economic, social and environmental factors that are *interdependent*.

Government intervention is used extensively to deal with macroeconomic issues. There are numerous policy options governments have at their disposal to address the economic challenges we studied in previous chapters.

Chapter 13 will critically examine many such policies that are put forward as possible means to deal with the problems of inflation, unemployment and economic growth.

Chapter 13

Demand-side and supply-side policies

BEFORE YOU START

- In what ways do governments generate revenue?
- What do governments spend money on and why do they prioritise certain choices over others?
- Can you think of ways the government intervenes to solve the problems of unemployment and inflation?

In this chapter we will use the *AD-AS* model as the basis for analysing and evaluating policy alternatives that can be used by governments to achieve a variety of macroeconomic objectives.

Photo: Washington D.C., USA. Federal Reserve, the US central bank

13.1 Introduction to macroeconomic policies

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- outline the basic principles of demand-side and supply-side policies (AO1)

Demand-side policies

Demand-side policies, also known as **demand management**, focus on changing aggregate demand, or shifting the aggregate demand curve in the *AD-AS* model, to achieve several macroeconomic goals. They are based on the idea that short-term fluctuations in real GDP of the business cycle are due to actions of firms and consumers affecting aggregate demand, causing inflationary or deflationary/recessionary gaps. Demand-side policies try to counteract the effects of these actions and bring aggregate demand to the full employment level of real GDP, or potential GDP.

There are two types of demand-side policies:

- monetary policy
- fiscal policy.

Monetary and fiscal policies attempt to reduce the short-run fluctuations of the business cycle. They are called *stabilisation policies*, because they try to 'stabilise' the economy, eliminating short-run instabilities caused by increases and decreases of aggregate demand. If stabilisation policies worked as intended, the business cycle would be flattened out, and the economy's actual output would be very close to its potential output (see Chapter 8, Figure 8.4). In practice, the most that stabilisation can hope to achieve is to lessen the severities of the business cycle.

Supply-side policies

Supply-side policies focus on the production and supply side of the economy, specifically on factors aimed at shifting the long-run aggregate supply (*LRAS*) or Keynesian *AS* curves to the right, to increase potential output and achieve long-term economic growth (see Chapter 11, Figure 11.2). They do not attempt to stabilise the economy by reducing the fluctuations of the business cycle. Instead, they focus on increasing the quantity and quality of factors of production, as well as on institutional changes intended to improve the economy's productive capacity.

There are two major categories of supply-side policies:

- market-based, which rely on the working of the market
- interventionist, which rely on government intervention.

13.2 Demand management and monetary policy

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the goals of monetary policy (AO2)
  - low and stable rate of inflation including inflation targeting
  - low unemployment
  - reduction of business cycle fluctuations
  - promotion of a stable economic environment for long term growth
  - external balance
- explain how equilibrium interest rates are determined (AO2) (HL only)
- draw a diagram to show determination of equilibrium interest rates (AO4) (HL only)
- explain how commercial banks create money (AO2) (HL only)
- explain the tools of monetary policy (AO2) (HL only)

CONTINUED

- open market operations
- minimum reserve requirements
- changes in central bank minimum lending rate
- quantitative easing

- distinguish between real and nominal interest rates and calculate real interest rates from data (AO2) (AO4)
- explain and evaluate expansionary and contractionary monetary policies to close inflationary and deflationary/recessionary gaps (AO3)
- draw *AD-AS* diagrams to show expansionary and contractionary monetary policy (AO4)
- discuss constraints on monetary policy including the limited scope of reducing interest rates when these are approaching zero, and consequences of low consumer and business confidence (AO3)
- discuss strengths of monetary policy including that it is incremental, flexible, easily reversible and has short time lags (AO3)
- evaluate monetary policy with respect to promoting low unemployment, low and stable rate of inflation and growth (AO3); see Section 13.7

The role of central banks

Monetary policy is carried out by the central bank of each country. The central bank must be distinguished from commercial banks. **Commercial banks** are financial institutions (which may be private or public) whose main functions are to hold deposits for their customers (consumers and firms), to make loans to their customers, to transfer funds by cheque (check) electronically from one bank to another, and to buy government bonds. (Government bonds were explained in Chapter 11 in the section on Government debt.)

The **central bank** is usually a government financial institution with several important responsibilities:

- **Banker to the government.** The central bank acts as a banker to the government in the way that commercial banks act as bankers to their customers. It holds the government's cash (as deposits), receives payments for the government and makes payments for the government, and manages the government's borrowing by selling bonds to commercial banks and the public.
- **Banker to commercial banks.** The central bank also acts as a banker to commercial banks by holding deposits for them and can also make loans to them in times of need. (It is not a banker to consumers and firms.)
- **Regulator of commercial banks.** The central bank regulates and supervises commercial banks, making sure they operate with appropriate levels of cash, according to rules that ensure the safety of the financial system.
- **Conduct monetary policy.** The central bank is responsible for monetary policy, based on its *control of the supply of money and interest rates*.

Every country has a central bank. In the countries of the European Union that have formed the European Monetary Union (that have adopted the euro, also known as 'euro zone' countries), the national central banks maintain many of their functions, noted above, but the responsibility for monetary policy has been transferred to a single organisation, the European Central Bank.

Central bank independence

Although the central bank is usually a government institution, in many countries it has a degree of independence from government interference in the pursuit of monetary policy. Independence ensures that monetary policy can be conducted in the best longer-term interests of the economy, without interference from political pressures (such as encouraging economic activity just before an election). There is a general trend around the world for governments to make central banks increasingly independent.

The goals of monetary policy

Monetary policy attempts to achieve the following goals:

- **Low and stable rate of inflation.** In Chapter 10 we learned that high rates of inflation have several undesirable effects on the economy and the population. Monetary policy attempts to achieve a low and stable rate inflation, which varies from country to country but is often about 2%; low but not so low that it is dangerously close to deflation (see Chapter 10). Many central banks practice the policy of *inflation targeting*, discussed below.

- **Low unemployment.** We have seen in Chapter 10 that unemployment also has several economic as well as personal and social costs. One of the goals of monetary policy may be to try to maintain unemployment at relatively low levels. The type of unemployment involved here is cyclical unemployment, arising in a deflationary gap due to insufficient aggregate demand.
- **Reduce business cycle fluctuations.** We studied the business cycle in Chapter 8. Figure 8.3 showed that real GDP growth is uneven and irregular. Fluctuations around potential output are disruptive to the normal functioning of the economy, causing inflation when output is above potential output, and cyclical unemployment when it is below the potential. One of the objectives of monetary policy is to try to make the fluctuations of the business cycle as small as possible.
- **Promote a stable economic environment for long-term growth.** Consumers and firms need a stable economic environment to be able to plan and carry out their economic activities. Firms, in particular, must make plans such as what capital goods to invest in, how much to invest, and whether, how and in what areas to pursue research and development (R&D) and technological innovations. To be able to plan, firms need economic stability, consisting of avoidance of sharp economic upturns (inflation and inflationary gaps) and downturns (recession and unemployment in deflationary gaps). Monetary policy helps create the macroeconomic environment that encourages activities impacting on long-term economic growth.
- **External balance.** External balance refers to a situation where a country's revenues from exports are balanced by spending on imports over an extended period of time. This is partly the result of the value of the country's currency, or its exchange rate (Chapter 16). The central bank can influence exchange rates because of the close relationship between interest rates and exchange rates.

Inflation targeting

While full employment and a low and stable rate of inflation are among the goals of monetary policy, in recent years more and more central banks around the world are using a kind of monetary policy that aims at maintaining a particular targeted rate of inflation (for example, Australia, Brazil, Canada, Chile, Finland, Israel, Mexico, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, the European Union and many others).

The International Monetary Fund (IMF) defines **inflation targeting** as:

> *'. . . the public announcement of medium-term numerical targets for inflation with an institutional commitment by the monetary authority to achieve these targets.'*[1]

Many countries pursuing inflation targeting have targets between 1.5% and 2.5%, with one percentage point above and below as a 'tolerance' margin. The inflation target is set in terms of the consumer price index (CPI), which also takes into account prices of imported goods (included in the CPI basket). However, inflation targeting is usually based on forecasts or predictions of *future* inflation based on the CPI (see Chapter 10 for a discussion of the CPI).

Inflation targeting offers a number of advantages including:

- achievement of a low and stable rate of inflation
- improved ability of economic decision-makers (firms, consumers) to anticipate the future rate of inflation and therefore plan their economic activities
- greater co-ordination between monetary and fiscal policy since knowledge about inflation targets allows the government to plan its fiscal policy to complement the central bank's monetary policy.

Yet there are also disadvantages:

- reduced ability of the central bank to pursue other macroeconomic objectives, particularly the goal of full employment; this is especially important in view of the potential conflict between a low rate of inflation and low unemployment (see Chapter 10)

[1] The IMF is an international financial institution that we will study in Chapter 20. IMF, 'De Facto Classification of Exchange Rate Regimes and Monetary Policy Frameworks', 31 April 2008 (http://www.imf.org/external/np/mfd/er/2008/eng/0408.htm).

- reduced ability of the central bank to respond to supply-side shocks; in the event of a supply-side shock, such as a sudden increase in oil prices leading to cost-push inflation and stagflation, the central bank may need flexibility to pursue an expansionary policy to bring the economy out of recession; this may mean a higher rate of inflation than the target
- an inflation target that is too low may lead to higher unemployment; if it is too high, it could lead to the problems resulting from high inflation.

Determination of the rate of interest (HL only)

The money market and the rate of interest

Monetary policy impacts indirectly on aggregate demand through the **interest rate** (or **rate of interest**). To understand monetary policy, we must consider how the rate of interest is determined.

When we borrow money, we must make a payment for the loan in addition to repaying the principal (the amount of the loan); this payment for a loan is **interest**. Interest is usually expressed as a percentage of the principal to be paid per year, called the *rate of interest*. If you borrow $1000 for one year at the rate of interest of 10% per year; at the end of the year you must pay back the principal of $1000, plus $100 of interest (10% of $1000).

In the real world there are many different rates of interest, depending on several factors, such as the level of risk of a loan (the greater the risk, the higher the interest rate); the amount of time over which the loan must be paid, known as 'maturity' (the longer the time period, the higher the interest rate); the size of the loan (the larger the loan, the lower the interest rate); the degree of market power of the lender (the greater the market power, the higher the interest rate), and others. However, when economists analyse the rate of interest in economic models (as we are doing here), they simplify the analysis by adopting the common practice of referring to 'the rate of interest' as if there were only one.

We can understand how the rate of interest is determined as an application of supply and demand in a special market, the money market, shown in Figure 13.1(a). **Money** is defined as anything that is acceptable as payment for goods and services; it includes currency (coins and paper money) and cheque (checking) accounts. In the money market the demand for money and the supply of money determine the equilibrium rate of interest. The horizontal axis measures the quantity of money in the economy, and the vertical axis measures the rate of interest.

The rate of interest can be thought of as the 'price' of money services. The **supply of money** is fixed at a level decided upon by the central bank. (We will see later in this chapter how this is done.) It appears in Figure 13.1(a) as a vertical line, S_m, because it does not depend on the rate of interest.

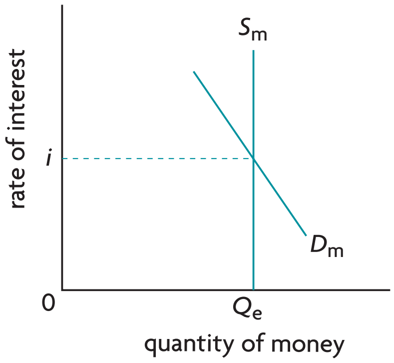
The *demand for money*, D_m, has the familiar downward-sloping shape of a demand curve. As the rate of interest falls, the quantity of money demanded by the public (consumers, firms, the government) increases. To see why, remember that money is defined to be currency and cheque (checking) accounts; the important thing to note here is that *money does not earn interest*. Suppose you have some savings; let's also say you have a choice between putting your savings in a form that earns interest, such as a savings deposit in a bank, or else you can hold your savings in the form of currency or a checking account in a bank that does not earn interest. Clearly, the higher the interest rate, the less attractive it is for you to hold money, and the lower the quantity of money you are likely to demand. This is the explanation behind the downward sloping demand for money curve.

The point of intersection between D_m and S_m determines the equilibrium rate of interest, i, illustrated in Figure 13.1(a).

If the central bank changes the money supply, the S_m curve shifts, determining a new rate of interest. This is shown in Figure 13.1(b). Suppose initially the money supply is at S_{m1}; with demand for money D_m, the equilibrium rate of interest is i_1. If the central bank increases the money supply, S_{m1} shifts to S_{m2}, and the equilibrium rate of interest falls to i_2. If the central bank decreases the money supply, S_{m1} shifts to S_{m3}, and the equilibrium rate of interest rises to i_3.

An increase in the supply of money leads to a fall in the rate of interest; a decrease in the supply of money leads to an increase in the rate of interest.

a Equilibrium rate of interest



b Changes in the supply of money cause changes in the equilibrium rate of interest

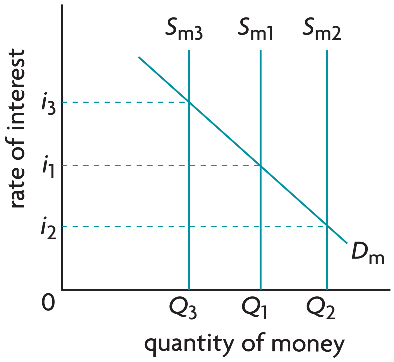


Figure 13.1: The money market and determination of the rate of interest

Setting a target interest rate

The central bank decides upon a target interest rate it wants to achieve, and then takes steps to adjust the money supply so that the actual equilibrium interest rate will become equal to the target interest rate.

To see how this works, suppose the Bank of Riverland (Riverland's central bank) decides to increase the interest rate from i_1 to i_3 in Figure 13.1(b). To do so, it takes measures to reduce the supply of money until the interest rate increases to i_3. If the actual market interest rate deviates from the target rate, it will continue to adjust the money supply in order to achieve the target rate. You can see, then, that central banks do not actually set or fix interest rates, but rather allow these to be determined by the market.

Therefore, if you hear in the news that the Bank of Riverland increased the interest rate from 3.25% to 3.50%, you would understand that 3.50% is the new target interest rate that the Bank of Riverland is trying to achieve by reducing the money supply.

In the real world there are many interest rates (as explained above), so what interest rate do central banks target? This varies from country to country, depending on the nature of the monetary system. In the United Kingdom, the central bank targets the 'base rate', which is the interest rate at which the Bank of England (the central bank) lends to commercial banks. In the United States, the Federal Reserve (the central bank) targets the 'federal funds rate', which is the rate used by commercial banks to borrow and lend from and to each other over a 24-hour period. The European Central Bank (of the euro zone countries) targets the 'minimum refinancing rate', which is the interest rate paid by commercial banks when they borrow from their respective national central bank to refinance their accounts.

TEST YOUR UNDERSTANDING 13.1

1 Distinguish between a country's central bank and commercial banks.

2 Outline the goals of monetary policy.

3 (HL only)

   a Using a diagram, explain how equilibrium interest rates are determined.

   b Outline what a central bank would do if it wanted **i** to lower interest rates, and **ii** to increase interest rates.

How the central bank changes the money supply (HL only)

To understand the process of **money creation** by commercial banks, we must first see how money is created. As we will now discover, money is created by commercial banks, though *how much* money they can create is determined by the central bank.

How commercial banks create money

Suppose you have earned £1000 in your summer job, and you would like to deposit this in your local commercial bank. You go to your bank and open a deposit, into which you place your £1000. What happens to your £1000? Your bank is unlikely to keep this money in its vaults for you to collect it when you wish. Instead, it is likely to lend out a good portion of it to other people who want to borrow from the bank.

In general when commercial banks receive deposits from their customers, they do not keep all this cash within their vaults. The funds they must legally keep are called *required reserves*, which are a legally determined fraction of total deposits, called the **minimum reserve requirement** or *required reserve ratio*. The rest are called *excess reserves* and can be lent out.

Suppose the minimum reserve requirement in your bank is 20%. The bank must keep £200 of required reserves in its vaults, and can lend out the remaining £800 of excess reserves. Suppose individual A borrows the £800 and buys a computer from individual B, who then deposits this amount in her bank. This bank must keep 20% of £800 or £160 (= 0.20 × £800), and can lend out the remaining amount, which is £640 (= £800 – £160). This process continues an infinite number of times, and in the end the amount of new loans that have been created will be £4000.

This amount is the result of a process where the amount of initial excess reserves of £800 are multiplied by a 'monetary multiplier', equal to $\frac{1}{\text{required reserve ratio}}$, which in this case is:

$$\frac{1}{0.20} = 5.$$

Therefore the amount of new loans that have been created = 5 × £800 = £4000. But *these new loans are none other than new money created*, since all the borrowers from the bank were able to use their loans to carry out their transactions by use of money.

Note that if the minimum reserve requirements had been lower, say 15%, the bank that received your initial deposit of £1000 would have excess reserves of £850 rather than £800, it would therefore be able to lend out this £850, the monetary multiplier would have been $\frac{1}{0.15} = 6.67$, so the amount of new loans, or actually new money, that could have been created would be = 6.67 × £850 = £5670.

Note also that this amount of new money is a *maximum amount* that can be created by a commercial bank given the initial deposit. It does not mean that the bank will actually make all those loans. In actual fact it is possible that the bank will create less than the maximum allowable amount.

When banks make loans, they are actually *creating new money*. The lower the minimum reserve requirement, the greater the excess reserves, the more loans can be made by commercial banks, and the more new money can be created. The minimum reserve requirement determines the maximum amount of new money that can be created.

This process of money creation is based on the idea that only a fraction of deposits need to be kept in the bank's vaults, it is therefore called a *fractional reserve system*.

Tools of monetary policy

We will now see how the central bank controls the amount of new money creation by examining the monetary policy tools of central banks.

Open market operations

The most important tool used by central banks to influence the supply of money is **open market operations**. Open market operations work by use of *bonds* (see Chapter 11 in connection with how governments borrow). Bonds are simply debt. The borrower issues a certificate called a bond that promises to pay interest at various intervals until a certain date when the money is repaid to the bond holder. The holder of the bond is therefore the lender, and the issuer of the bond is the borrower. Open market operations work through the buying and selling of *pre-existing* government bonds in the bond market. The bond market is simply a market where holders of bonds can buy and sell *pre-existing* bonds.

Suppose a central bank wishes to lower the interest rate, and must therefore increase the money supply. It will *buy* government bonds from commercial banks. When the central bank buys the bonds it pays the commercial banks for these. This process increases the commercial banks' excess reserves, which they can use to make more loans, and therefore the money supply increases giving rise to lower interest rates, as Figure 13.1 shows.

On the other hand if the central bank wants to raise the interest rate, it *sells* bonds to commercial banks; as the banks must pay the central bank for these, their excess reserves and therefore their lending ability are reduced, and the money supply is lowered. In Figure 13.1 the money supply curve shifts to the left, resulting in a higher interest rate.

Central banks may buy and sell bonds not only from commercial banks but also from the public in general. The end result is the same, only it occurs indirectly.

Minimum reserve requirements

This tool involves changes of the minimum reserve requirements by the central bank. As we saw earlier, if the reserve requirements decrease, this means that the commercial banks excess reserves increase, therefore their lending ability increases, so too their ability to create money. Hence the money supply increases. On the other hand if the reserve requirements increase, the result is the opposite. Excess reserves drop, the banks' lending ability decreases, so too their ability to create money, hence the money supply decreases.

Changes in the central bank's minimum lending rate

One of the functions of a central bank is that it sometimes lends to commercial banks. When it does this it charges them an interest rate, known as a **minimum lending rate** (according to UK terminology). This interest rate has different names in different countries, for example in the European Union is it known as the *refinancing rate*, in the United States as the *discount rate*, in the United Kingdom as the *base rate*.

If commercial banks want reserves to increase their lending they can borrow from the central bank. Therefore, the minimum lending rate reflects the cost to commercial banks of acquiring more reserves. If the central bank decreases this interest rate, it becomes less costly for commercial banks to borrow from the central bank, and so they can increase their borrowing, increase their reserves, therefore increasing the money supply. If the central bank increases this rate, borrowing becomes more costly for the commercial banks, therefore their lending ability is reduced and the money supply decreases.

Quantitative easing

Quantitative easing is similar to buying bonds in open market operations, but on a much larger scale, involving more types of financial assets and larger quantities of these. It is an unconventional type of monetary policy that was first used by Japan in 2001, and later by other central banks due to the global financial crisis. The United States began using the policy in 2008 and the European Central Bank in 2015. Conventional monetary policy had reduced interest rates to very low levels, approaching zero. The objective was to encourage borrowing by firms and consumers in order to increase aggregate demand. But when interest rates fall so low conventional monetary policy becomes ineffective (we will see why later in this chapter). With quantitative easing, the central bank buys huge quantities of assets that commercial banks have or own. In order to pay for the assets the central bank creates reserves electronically for the commercial banks. As a result the commercial banks that sell the assets end up with many more reserves which they can then use to make loans, the objective being to increase aggregate demand.

Table 13.1 summarizes the tools of monetary policy.

	To lower the interest rate the central bank will	To increase the interest rate the central bank will
Open market operations	Buy bonds increasing commercial bank reserves thus increasing the money supply	Sell bonds decreasing commercial bank reserves thus decreasing the money supply
Minimum reserve requirements	Lower reserve requirements increasing commercial bank reserves thus increasing the money supply	Increase reserve requirements decreasing commercial bank reserves thus decreasing the money supply
Central bank minimum lending rate	Lower minimum lending rate increasing commercial bank reserves thus increasing the money supply	Increase minimum lending rate decreasing commercial bank reserves thus decreasing the money supply
Quantitative easing	Create new reserves electronically used by the central bank to buy a huge variety and quantity of assets thus directly increasing the money supply	

Table 13.1: Tools of monetary policy

TEST YOUR UNDERSTANDING 13.2 (HL ONLY)

1 Using the concept of minimum reserve requirements, explain the process by which commercial banks create money.

2 Explain how a central bank could use the tools of

a open market operations,

b minimum reserve requirements,

c changing the minimum lending rate to i increase the supply of money, ii to decrease the supply of money.

3 Explain what is meant by the policy of quantitative easing.

Real versus nominal interest rates

The **nominal rate of interest** is simply the market rate that prevails at any moment in time. If your bank tells you that you will receive 5% interest on your savings, that is the nominal interest rate. The **real rate of interest** is the interest rate that has been corrected for inflation. When we know the rate of inflation and the nominal interest rate, we can calculate the real interest rate

real interest rate = nominal interest rate – rate of inflation

If the annual rate of inflation is 3% and the nominal interest rate is 5% per year, the real rate of interest is 5% – 3% = 2% per year. So if you have \$1000 in a bank account that earns 5% per year, after one year this will be a nominal amount of \$1050 (= 1000 × 1.05) but in real terms, or in terms of the purchasing power of your original \$1000 you will have \$1020 (= 1000 × 1.02).

When we studied the costs of inflation, we learned that savers can protect themselves against losses in the real value of their saving if they can receive an interest rate that is at least as high as the rate of inflation. If the annual rate of inflation is 3% but the nominal interest rate is only 2% per year, the real rate of interest will be 2% – 3% = –1%. Your \$1000 in real terms will be worth \$990 (= 1000 × 0.99) after one year. In order to protect yourself against inflation you must receive a nominal rate of interest that is at least 3%, equal to the rate of inflation.

The role of monetary policy: deflationary/recessionary and inflationary gaps

Changes in interest rates and aggregate demand

The point of changing the money supply so as to change interest rates is ultimately to influence aggregate demand. Changes in interest rates affect two of the four components of aggregate demand: investment, *I*, and consumption, *C* (see Chapter 9, Table 9.1). Since some consumer and firm spending is paid for by borrowing, a change in interest rates is intended to affect the amount of consumer spending (*C*) and investment spending (*I*).

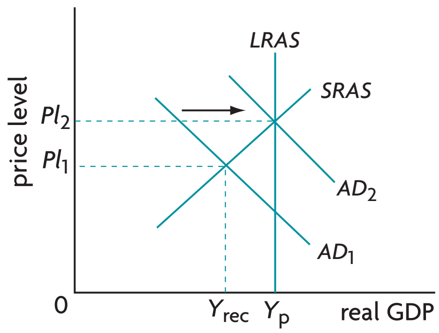
An increase in interest rates is intended to lower consumer and business borrowing and hence spending (lower *C* and *I*), and therefore shift *AD* to the left. A decrease in interest rates is intended to increase consumer and business borrowing and hence spending (higher *C* and *I*), and therefore shift *AD* to the right.

Expansionary (easy) monetary policy

Suppose the economy is experiencing a deflationary (recessionary) gap due to insufficient aggregate demand, as in Figure 13.2(a) and (b). The central bank decides to increase the money supply, causing a rightward shift in the supply of money curve from S_{m1} to S_{m2} in Figure 13.1(b). With the demand for money constant, the interest rate falls from i_1 to i_2.

The drop in the interest rate means a lower cost of borrowing; therefore, consumers and firms are likely to borrow more and spend more, so that consumption spending (*C*) and investment spending (*I*) increase. The effect is to increase aggregate demand and cause a rightward shift of the *AD* curve. This is shown in Figure 13.2(a) and (b), where the recessionary gap has been closed through the shift from AD_1 to AD_2.

Both the monetarist/new classical and the Keynesian models predict that an increase in *AD* increases real GDP. However, the size of the increase in real GDP will not be the same. It will be smaller in the monetarist/new classical model than in the Keynesian one, because of the upward-sloping *SRAS* curve. The effects differ also in the case of the price level. In the monetarist/new classical model, the increase in *AD* always results in a rise in the price level because of the upward-sloping *SRAS* curve. In the Keynesian model, the increase in *AD* may result in no increase in the price level at all if the *AD* shift occurs a The monetarist/new classical model



b The Keynesian model

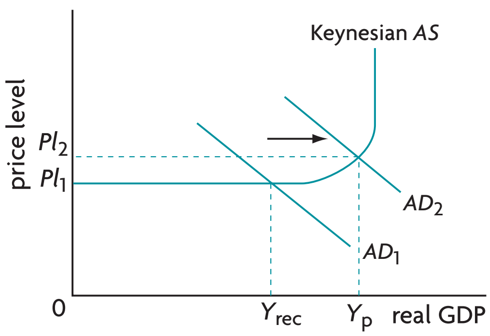


Figure 13.2: Effects of expansionary policy: eliminating a recessionary/deflationary gap

entirely within the horizontal section of the *AS* curve. If the *AD* shift reaches into the upward-sloping part of the Keynesian *AS* curve, as in Figure 13.2(b), there will be only a smaller increase in the price level.

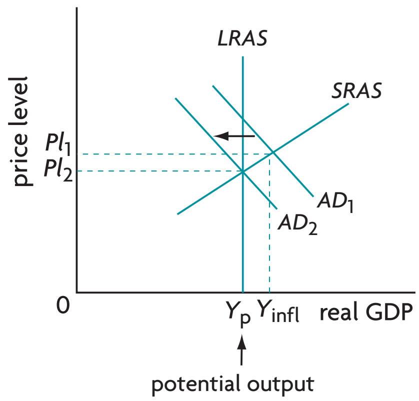
An increase in the money supply by the central bank is referred to as an **expansionary monetary policy** since the objective is to expand aggregate demand and the level of economic activity. It is also an *easy money policy*, since it results from an increase in the supply of money compared to the monetarist/new classical model.

Contractionary (tight) monetary policy

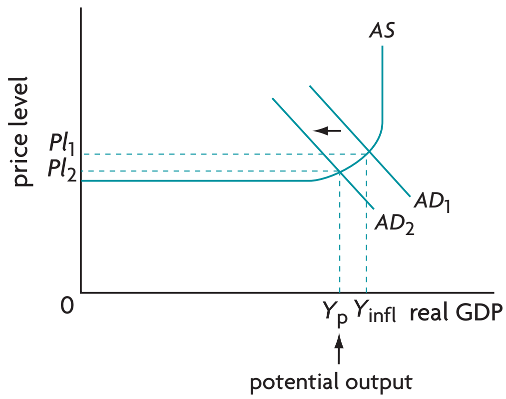
Suppose now that the economy is experiencing an inflationary gap caused by excess aggregate demand, as in Figure 13.3(a), (b) and (c) where the aggregate demand curve AD_1 intersects the *SRAS* curve and the Keynesian *AS* curve at a level of real GDP, Y_{infl}, that is greater than the full employment or potential output level Y_p. The central bank reduces the money supply; this appears in Figure 13.1(b) as a leftward shift of the S_m curve, from S_{m1} to S_{m3}. With the demand for money constant, the result is a higher rate of interest, i_3, or a higher cost of borrowing, and therefore reduced borrowing by consumers and firms. The effect of lower investment spending (*I*) and lower consumer spending (*C*) is to decrease aggregate demand. This is shown in all parts of Figure 13.3, where the inflationary gap has been closed through the shift from AD_1 to AD_2.

A decrease in the money supply by the central bank is referred to as a **contractionary monetary policy**, as the objective is to contract aggregate demand and therefore the economy. It is also known as a *tight money policy*, in view of the decrease in the supply of money.

a The monetarist/new classical model



b The Keynesian model



c The Keynesian model with ratchet effect

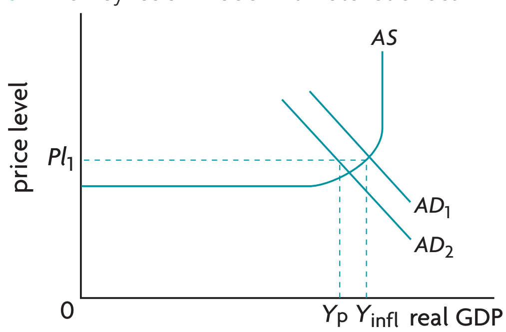


Figure 13.3: Effects of contractionary policy: eliminating an inflationary gap

Note that the effects of a fall in aggregate demand may be different depending on the model considered. If *AD* falls within the upward-sloping part of the *AS* curve in the Keynesian model (as in part (b)) the effects on the price level and real GDP are similar in the two models (parts (a) and (b)) because the slope of the curves is similar. But if *AD* were to decrease into the horizontal part of the *AS* curve, there would be a larger fall in real GDP and a smaller fall in the price level in the Keynesian model, or none at all, compared with the monetarist/new classical model. (The argument is analogous to that noted above in connection with expansionary policy.)

The Keynesian model with the ratchet effect

We know that a feature of the Keynesian model is that the price level can easily increase with strong aggregate demand, but does not easily fall as aggregate demand decreases. Therefore, the decrease in the price level shown in Figure 13.3(b) does not make sense. To take this into account, many economists refer to the 'ratchet effect', shown in Figure 13.3(c).[2] According to the ratchet effect, the price level moves up when there is an increase in *AD*, and then remains at the same level until there is a further increase in *AD*. In Figure 13.3(c), the change from AD_1 to AD_2 causes real GDP to fall to Y_p, but the price level remains constant at pl_1. This is a more realistic representation of what usually happens in the real world. (You may note that the decrease in *AD* required to bring real GDP to Y_p from Y_{infl} is smaller with the ratchet effect than without.)

Monetary policy is carried out by the central bank, which aims at changing interest rates to influence the *I* and *C* components of aggregate demand. In a deflationary recessionary gap, the central bank may pursue an expansionary (easy money) policy through lower interest rates to encourage *I* and *C* spending, the objective being to shift the *AD* curve to the right leading to equilibrium at the full employment level of real GDP (potential GDP). In an inflationary gap, the central bank can pursue a contractionary (tight money) policy through higher interest rates aimed at discouraging *I* and *C* spending, causing the *AD* curve to shift to the left leading to equilibrium at the full employment level of real GDP (potential GDP).

TEST YOUR UNDERSTANDING 13.3

1 a Distinguish between nominal and real interest rates.

b Calculate the real interest rate if the nominal rate is 5% and the rate of inflation is 3%.

c Calculate the real interest rate if the nominal rate is 4% and the rate of inflation is 7%.

d Outline the meaning of a negative real interest rate.

2 a Distinguish between expansionary and contractionary monetary policy.

b State the components of aggregate demand that monetary policy can influence.

c Explain the role of the rate of interest in monetary policy.

3 Using diagrams, show how the government can use monetary policy when there is

a a recessionary/deflationary gap, and

b an inflationary gap.

4 Using the monetarist/new classical model, explain the impact on real GDP, the price level and unemployment, of the following policies of a country's central bank:

a a fall in the rate of interest

b an increase in the rate of interest.

5 a Answer parts (a) and (b) in question 3 above using the Keynesian *AD-AS* model.

b Explain how the predictions of the two models differ.

c Describe when the ratchet effect comes into play.

[2] A ratchet is a simple machine that allows for something to move only in one direction.

Evaluating monetary policy

Whereas monetary policy is intended to achieve particular objectives, it does not always work as expected.

Constraints on monetary policy

- **Possible ineffectiveness in recession.** Whereas monetary policy can work effectively when it raises interest rates to fight inflation, it is less certain to be as effective in a deep recession, because:
  - ***Interest rates cannot fall when approaching zero.*** As interest rates approach zero, they cannot fall further to encourage spending by firms and consumers.
  - ***Low consumer and producer confidence.*** If firms and consumers are pessimistic about future economic conditions, they may avoid taking out new loans, and may even reduce their investment and consumer spending, so that aggregate demand will not increase (it may even decrease).
  - ***Banks may be fearful of lending.*** In a severe recession, banks may be unwilling to increase their lending, because they may fear that borrowers might be unable to repay the loans.

  Such policy ineffectiveness is not something that happens often; however, it appears to have occurred during the Great Depression of the 1930s, in Japan in the late 1990s and early 2000s, and in the global recession that began in the autumn of 2008.
- **Conflict between government objectives.** Manipulation of interest rates affects not only variables in the domestic economy (consumption and investment spending, inflation, unemployment) but also variables in the foreign sector of the economy, such as exchange rates. The pursuit of domestic objectives may conflict with the pursuit of the goal of external balance in the foreign sector (see Chapters 16, 17).
- **May be inflationary.** If it lasts too long it may be inflationary, if aggregate demand increases beyond what is necessary to eliminate a deflationary/recessionary gap.
- **Problematic when dealing with stagflation or cost-push inflation**. Monetary policy is a *demand-side policy*, and is therefore unable to deal effectively with supply-side causes of instability.

Strengths of monetary policy

- **Interest rate changes can be incremental.** Interest rates can be adjusted in very small steps, making monetary policy well suited to 'fine tuning' of the economy.
- **Interest rates changes are reversible.** Interest rate changes can also be easily reversed if necessary. An expansionary policy can easily be reversed into a contractionary policy and vice versa.
- **Monetary policy is flexible.** Interest rates can be changed often according to needs.
- **Relatively short time lags (time delays).** While monetary policy can be implemented relatively quickly, it is subject to time lags as it takes time for interest rate changes to affect the economy, though these are not as long as in the case of fiscal policy.
- **Central bank independence.** Independence from the government discussed above means the central bank can take decisions that are in the best longer-term interests of the economy, and can therefore pursue policies that may be politically unpopular (such as higher interest rates making borrowing more costly).
- **Limited political constraints.** Monetary policy does not face political pressures as fiscal policy does, since it does not involve making changes in the government budget, whether in terms of government spending that would affect merit and public goods provision or government taxes (see the discussion on fiscal policy below).
- **No budget deficits or debt.** It does not lead to budget deficits or increased levels of debt as fiscal policy does in the case of expansionary policy.
- **No crowding out.** Monetary policy does not lead to crowding out, which may be a weakness of expansionary fiscal policy (this will be discussed below at HL).

TEST YOUR UNDERSTANDING 13.4

1. Explain why monetary policy is not very well suited to dealing with instabilities caused by decreases in SRAS.
2. Examine the strengths and weaknesses of monetary policy and discuss which of these you think are more important.

13.3 Demand management and fiscal policy

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- fiscal policy – distinguish between (AO2)
    - sources of revenue – direct and indirect taxation, sales of goods and services by state-owned enterprises, sale of government assets
    - expenditures – current expenditures, capital expenditures, transfer payments
- explain the goals of fiscal policy (AO2)
    - low and stable rate of inflation
    - low unemployment
    - reduction of business cycle fluctuations
    - promotion of a stable economic environment for long term growth
    - equitable distribution of income
    - external balance
- explain and evaluate expansionary and contractionary fiscal policies to close inflationary and deflationary/recessionary gaps (AO3)
- draw *AD-AS* diagrams to show expansionary and contractionary monetary policies (AO4)
- discuss constraints on fiscal policy including political pressures, time lags, sustainable debt (AO3)
- discuss the constraint of crowding out and draw a diagram illustrating it (HL only) (AO3) (AO4)
- discuss strengths of fiscal policy including the ability to target specific sectors, effectiveness of government spending in deep recession (AO3)
- discuss automatic stabilisers (unemployment benefits, progressive taxes) (HL only) (AO3)
- evaluate fiscal policy with respect to promoting low unemployment, low and stable rate of inflation and growth (AO3); see Section 13.7

The government budget

The government budget is a type of plan of a country's revenues and expenditures over a period of time (usually a year) that the government makes to plan its activities.

Sources of government revenue

Government revenue consists of all the funds that flow toward the government from outsiders. The main sources of revenue are:

- **Taxes of all types**, both direct and indirect. Taxes are the most important source of government revenues.
- **From the sale of goods and services.** Governments provide many goods and services free of charge to their users (such as public goods). However there are many services for which the users must make a payment, including transportation, electricity, water and many more. The revenues from these sales usually go toward covering the government's costs of providing them.
- **From the sale of government-owned (state-owned) assets, or property.** Such sales are known as privatisation, which involves the transfer of ownership from the government to private owners.

Types of government expenditure

Governments have three types of expenditures:

- **Current expenditures** include the government's spending on day-to-day items that are recurring (repeat themselves) and items that are used up or 'consumed' as a good or service is provided. They include wages and salaries (for all government employees); spending for supplies and equipment for the day-to-day operation of government activities, including for example, school supplies and medical supplies for public schools and public health care services; provision of subsidies; and interest payments on government loans.

- **Capital expenditures** include public investments, or spending to produce physical capital, such as building roads, airports, harbours, school buildings, hospitals, etc.
- **Transfer payments** include payments by the government to vulnerable groups for the purposes of income redistribution (for example, unemployment benefits, child allowances, etc.).

Both current and capital expenditures are included in the measurement of GDP; they are included under *G*, for 'government spending', in the expenditure approach to measuring GDP (see Chapter 8). However transfer payments are not, because they represent income that is redistributed away from taxpayers and toward the receivers of the transfer payments; they do not represent value of new output produced.

As we know from Chapter 11 if tax revenues are equal to government expenditures over a period of time the government is said to have a *balanced budget*. If expenditures are larger than tax revenues, there is a *budget deficit*; if expenditures are smaller than tax revenues, there is a *budget surplus*. When there is a budget deficit, the government finances (pays for) the excess of expenditures over revenues by borrowing. Over time, the government's accumulation of deficits minus surpluses is referred to as or *government debt* or *national debt* or *public debt*.

The goals of fiscal policy

Fiscal policy refers to manipulations by the government of its own expenditures and taxes to influence the level of aggregate demand. The goals of fiscal policy are the same as the goals of monetary policy, except that fiscal policy also has the goal of achieving an equitable distribution of income.

- **Low and stable rate of inflation** Fiscal policy tries to maintain a slow and stable rate of inflation by manipulating taxes and government spending to influence aggregate demand. Unlike monetary policy, fiscal policy does not have any inflation target, but may be used to complement monetary policy if monetary policy is not as effective as expected.
- **Low unemployment** Fiscal policy similarly may try to influence aggregate demand and therefore unemployment. Here too the type of unemployment involved is cyclical, when the economy is in a deflationary gap due to insufficient aggregate demand.
- **Reduce business cycle fluctuations** Fiscal policy tries to reduce the size of the fluctuations of the business cycle to make inflationary and deflationary gaps as small as possible.
- **Promote a stable economic environment for long-term growth** As with monetary policy, firms in particular need a stable economic environment without sharp inflationary or deflationary gaps in order to promote business confidence so that firms can carry out the activities needed for long-term economic growth.
- **External balance** Fiscal policy can help achieve external balance (where a country's revenues from exports are roughly equal to its spending on imports) by influencing the level of imports through its effects on aggregate demand. This topic will be presented in Chapter 17 (at HL).
- **Equitable distribution of income** This goal is not shared with monetary policy. Fiscal policy has major effects on the distribution of income by determining tax policies and government spending to produce and provide particular goods and services, such as merit goods. This topic was discussed in Chapter 12.

TEST YOUR UNDERSTANDING 13.5

1 Explain
   a the sources of government revenues, and
   b the types of government expenditures.

2 Outline
   a the goals of fiscal policy, and
   b the difference between 'government budget deficit' and 'government debt'.

The role of fiscal policy in demand management

We have seen that fiscal policy involves manipulations by the government of its own spending and taxes to influence aggregate demand. The components of aggregate demand are consumption (C), investment (I), government spending (G), and net exports ($X - M$). Fiscal policy can affect three of these components:

- The level of the government's own spending, G.
- The level of consumption spending, C, can be influenced if the government changes taxes on consumers (personal income taxes), changing their level of disposable income (consumer income after income taxes have been paid).
- The level of investment spending, I, can also be influenced if the government changes taxes on business profits.

Expansionary fiscal policy

The diagrams needed to understand fiscal policy are the same as those for monetary policy. Suppose the economy is experiencing a recessionary/deflationary gap caused by insufficient aggregate demand. In Figure 13.2(a) and (b) AD_1 intersects both the $SRAS$ curve and the Keynesian AS curve at a level of real GDP, Y_{rec}, that is below the full employment (potential output) level, Y_p. The effects of fiscal policy can be illustrated equally well by both the monetarist/new classical model in part (a) and the Keynesian model in part (b). The government's objective is to shift AD_1 to AD_2, where the economy will achieve full employment or potential output, Y_p, thereby closing the recessionary gap. This is **expansionary fiscal policy**, because it works to expand aggregate demand and the level of economic activity. Expansionary fiscal policy may consist of:

- increasing government spending
- decreasing personal income taxes
- decreasing business taxes (taxes on profits), or
- a combination of the above.

An increase in government spending impacts directly on aggregate demand, which increases. If the government decreases taxes, aggregate demand is affected in a two-step process. If personal income taxes are cut, there is a rise in disposable income, which is then likely to lead to an increase in consumption spending, causing the AD curve to shift to the right. If business taxes are cut, after-tax business profits increase, which in turn is likely to lead to higher investment spending and therefore higher AD. In all three cases, AD is intended to shift to the right from AD_1 to AD_2, allowing the economy to achieve full employment or potential output Y_p.

Finally, the government may decide to pursue a policy of increasing government spending and lowering taxes simultaneously. How can it increase its own spending while keeping taxes constant or decreasing them? It can do so by borrowing. If initially it has a balanced budget, an increase in G while taxes remain constant or fall creates a budget deficit. If it already has a budget deficit at the outset, the deficit will become larger. If it has a budget surplus at the outset, then the surplus either will become smaller, or it will turn into a deficit.

Note that the effects of the AD increase are different depending on the shape of the AS curves. In the monetarist/new classical model, there is a smaller increase in real GDP and a larger increase in the price level compared to the Keynesian model. This is analogous to the AD increase in the case of monetary policy.

Contractionary fiscal policy

Suppose the economy is experiencing an inflationary gap caused by excessive aggregate demand, shown in Figure 13.3(a) and (b). The government's objective now is to attempt to shift AD_1 to AD_2, so that AD_2 intersects aggregate supply at the full employment level of output, Y_p, thereby closing the inflationary gap. This is called **contractionary fiscal policy**, because it works to contract aggregate demand and the level of economic activity. Contractionary fiscal policy consists of:

- decreasing government spending
- increasing personal income taxes
- increasing business taxes (taxes on profits), or
- a combination of the above.

A decrease in government spending has a direct influence on the aggregate demand curve, causing it to shift to the left. An increase in personal income taxes or business taxes is intended to affect aggregate demand in a two-step process. As personal income taxes increase, after-tax income falls, causing consumption spending and aggregate demand to fall. As taxes on profits increase, after-tax profits fall, leading businesses to spend less on investment and causing aggregate demand to fall. In all three cases, the aggregate demand curve is meant to shift to the left.

The government can also pursue a combination of decreases in government spending with increases in personal income and business taxes. Depending on the initial conditions in the government's budget, such a combination of policies would lead to the creation of a budget surplus, or the shrinkage of a budget deficit, or turning a budget deficit into a surplus.

As in the case of monetary policy, if *AD* falls within the upward-sloping part of the *AS* curve in the Keynesian model, the effects on the price level and real GDP are similar to those in the monetarist/new classical model, shown in parts (a) and (b). But if *AD* were to decrease into the horizontal part of the *AS* curve, there would be a larger fall in real GDP and a smaller fall in the price level in the Keynesian model. Note also the ratchet effect, shown in Figure 13.3(c), which takes into account the unlikelihood of the price level falling in the Keynesian model as *AD* decreases.

Fiscal policy involves manipulations by the government of its own expenditures and taxes to influence the *G*, *C* or *I* components of aggregate demand. Expansionary fiscal policy can be used when there is a recessionary gap, and aims to shift the *AD* curve to the right leading to equilibrium at the full employment level of real GDP (potential GDP). Contractionary fiscal policy can be used when there is an inflationary gap, and aims to shift the *AD* curve to the left leading to equilibrium at the full employment level of real GDP (potential GDP).

The effects of both types of policy have been illustrated by use of the same diagrams (Figures 13.2 and 13.3). Yet this simple diagrammatical analysis hides important differences between the two types of policy, related to the different channels that affect spending of the various *AD* components. These differences are summarised in Table 13.2.

Expansionary policy in recessionary/deflationary gaps (recession)

Type of policy	Measures	Effects
Monetary policy	increase supply of money → lower interest rate →	
	(i) increase consumption spending	increase *AD*
	(ii) increase investment spending	increase *AD*
Fiscal policy	increase government spending	increase *AD*
	lower personal income taxes → increase consumption spending	increase *AD*
	lower business taxes (on profits) → increase investment spending	increase *AD*

Contractionary policy in inflationary gaps (inflation)

Type of policy	Measures	Effects
Monetary policy	decrease supply of money → raise interest rate →	
	(i) decrease consumption spending	decrease *AD*
	(ii) decrease investment spending	decrease *AD*
Fiscal policy	decrease government spending	decrease *AD*
	raise personal income taxes → decrease consumption spending	decrease *AD*
	raise business taxes (on profits) → decrease investment spending	decrease *AD*

Table 13.2: Demand-side policies to correct deflationary/recessionary and inflationary gaps

TEST YOUR UNDERSTANDING 13.6

1 a Distinguish between expansionary and contractionary fiscal policy.

b Identify what components of aggregate demand are affected by fiscal policy; outline how they are affected.

2 Using diagrams, show how the government can use fiscal policy when there is

a a recessionary gap, and

b an inflationary gap.

3 Using the monetarist/new classical model, explain how the following policies can impact on real GDP, the price level and unemployment:

a The government lowers income taxes.

b The government decreases its spending on defence.

c The government increases taxes on business profits.

d The government increases its spending on the country's road and highway system.

4 a Answer parts (a)–(d) in question 4 above using the Keynesian *AD-AS* model.

b Explain how the predictions of the two models differ.

c State when the ratchet effect comes into play.

Evaluating fiscal policy

Fiscal policies also do not always achieve the desired and expected impacts, though mostly for different reasons than monetary policy.

Constraints on fiscal policy

- **Problems of time lags.** There are a number of time delays until:
  - the problem (recessionary or inflationary gap) is recognised by the government authorities and economists
  - the appropriate policy to deal with the problem is decided upon
  - the policy takes effect in the economy.

  Some months may pass in the case of each of these. By the time the policy action has taken effect the problem may have become less or more severe, so that the policy action is no longer the most appropriate one.
- **Political constraints** Government spending and taxation face numerous political pressures. Spending for social services (merit goods) and public goods cannot easily be cut if a contractionary policy is required. On the other hand, tax increases are politically unpopular and may be avoided by the government even though they might be necessary. Tax decreases could also be inappropriately enacted because they are politically popular. Therefore political factors may sometimes lead to unsuitable fiscal policies.
- **Sustainable debt** As explained above, deficits arise when government spending is more than government revenues, financed by borrowing, which contributes to the building up of debt. Sustainable debt refers to a level of debt where a borrowing government can meet its present and future debt obligations (interest payments plus repayment of capital) without accumulating overdue debt payments. Particularly if the economy is in recession, when tax revenues fall (as unemployment rises) and government spending increases (on unemployment benefits) deficits are likely to increase. Over an extended period these may create a problem of unsustainable debt possibly leading to default (inability to pay back debts).
- **In a recession, tax cuts may not be very effective in increasing aggregate demand** Tax cuts are less effective in a recession than increases in government spending because part of the increase in after-tax income is saved. If the proportion of income saved rises due to pessimism about the future, the impacts of tax cuts on aggregate demand are even weaker. Increases in government spending are more powerful because they work in their entirety to increase aggregate demand.
- **Inability to 'fine tune' the economy** Whereas fiscal policy can lead the economy in a general direction of larger or smaller aggregate demand, it cannot 'fine tune' the economy; it cannot be used to reach a precise target with respect to the level of output, employment and the price level.
- **May be inflationary** If it lasts too long it may be inflationary, this may occur if aggregate demand increases beyond what is necessary to eliminate a deflationary/recessionary gap. This weakness is similar to monetary policy.

- **Inability to deal with cost push inflation, or stagflation** Fiscal policy, like monetary policy is a *demand-side policy*, and is therefore unable to deal effectively with supply-side causes of instability.
- **Crowding out (HL only)** If the government pursues an expansionary fiscal policy involving spending increases without an increase in revenues, it is forced to borrow. Government borrowing involves an increase in the demand for money, which leads to an increase in the rate of interest. A higher interest rate in turn can lead to lower investment spending by private firms, or a 'crowding out' of private investment. This means that the government's expansionary fiscal policy is weakened, since a greater *G* (government spending) is counteracted by a lower *I* (investment spending). Crowding out is illustrated in Figure 13.4. In part (a) there is a rightward shift from AD_1 to AD_2 due to the increase in *G*, and a leftward shift from AD_2 to AD_3 due to the fall in *I*. This shows partial crowding out, where the fall in investment spending is smaller than the increase in government spending. Part (b) shows complete crowding out, where the fall in *I* is equal to the increase in *G*. Crowding out is controversial. Some economists, mainly in the Keynesian tradition, believe that in a recession, the stimulus provided to the economy by the government's increased spending may raise output and employment, improve business expectations about their future sales, and increase investment spending in spite of the increase in the interest rate. In this case, the government's deficit spending is less likely to crowd out private investment. Other economists, mainly in the monetarist/new classical tradition, believe that investment spending will be crowded out in the event of deficit financing even in a recession.

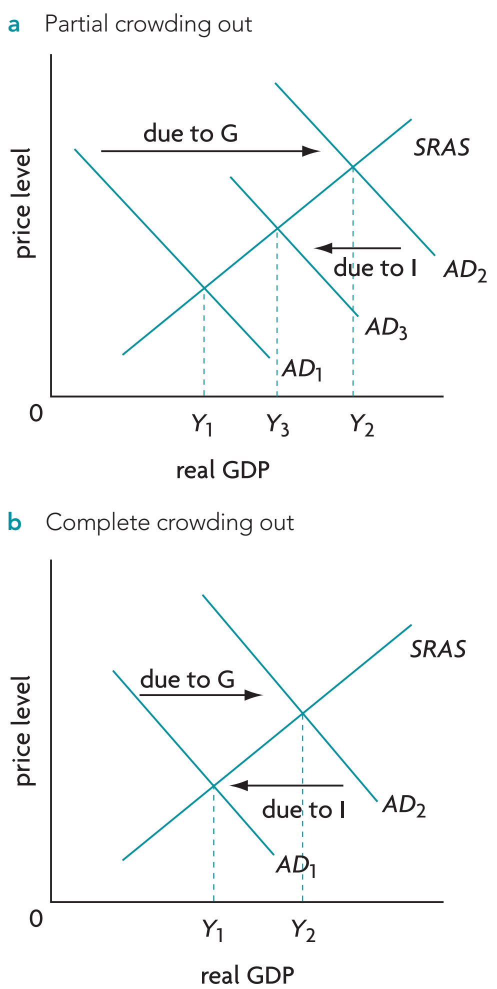


Figure 13.4: Crowding out of private investment

Strengths of fiscal policy

- **Pulling an economy out of a deep recession** Until the Great Depression of the 1930s, classical economists believed that short-term economic fluctuations were self-correcting: in a recession, wage and price flexibility would make the economy return to full employment. (This is like the thinking of monetarist/new classical economists; see Chapter 9.) Yet the experience of the Great Depression, with low levels of output and incomes and high unemployment over a long period of time, showed that market forces alone were unable to pull the economy out of the deep recession. In the now classic work, *The General Theory of Employment, Interest and Money* (1936), John Maynard Keynes (the originator of 'Keynesian economics') argued that wages and prices were inflexible in the downward direction even in the face of steep recession, and that low aggregate demand could keep the economy stuck in a recessionary gap indefinitely. This can occur when the *AD* curve intersects the Keynesian aggregate supply (*AS*) curve at a point on its horizontal section, where real GDP is less than full employment GDP. The strength of fiscal policy is to pull an economy out of a deep recession. In the global recession that began in autumn 2008, fears of a major global recession made governments around the world turn to expansionary fiscal policy to stimulate low aggregate demand.
- **Ability to target sectors of the economy** Fiscal policy can target spending in specific sectors according to

government priorities. For example, it may focus on changing the amount of spending on education and particular levels within education; health care, focusing on particular social groups; infrastructure and particular types of infrastructure (airports, roads, hospitals) or the locations of infrastructure, focusing if necessary on economically depressed regions; a variety of public goods (police force, public parks, etc.); and so on.

- **Direct impact of government spending on aggregate demand** Changes in government spending impact directly on aggregate demand, and this can be helpful to policy-makers who want to be reasonably certain that changes in spending are likely to change aggregate demand in the desired direction. Changes in taxes are less direct, as they work by changing consumer disposable income and firm after-tax profits and this poses some uncertainties about their effects on aggregate demand.
- **Dealing with rapid and escalating inflation.** Inflationary pressures arising when there is an inflationary gap can sometimes get out of hand, resulting in rapid increases in the price level. Contractionary fiscal policy may then be used effectively to help bring the problem under control.
- **Ability to affect potential output** Fiscal policy can affect potential output and long-term economic growth indirectly (by creating a stable macroeconomic environment) and directly through investments in human capital and physical capital (infrastructure) and through offering incentives to firms to invest
- **Automatic stabilisers (HL only)** Automatic stabilisers are factors that automatically, without any action by government authorities, work toward stabilising the economy by reducing short-term fluctuations of the business cycle. There are two important stabilisers: progressive income taxes and unemployment benefits.
    - ***Progressive income taxes***. Income taxes are *progressive* when the fraction of income that is taxed increases as income increases (see Chapter 12). In the upswing of the business cycle, as real GDP and incomes rise, income taxes rise proportionately more than the rise in income, causing after-tax (disposable) income to be lower than it would otherwise be. This means aggregate demand increases less, and this counteracts the economic expansion, making it smaller than it would otherwise be.

      In a recession, the opposite occurs. With real GDP and incomes falling, income taxes fall proportionately more in a progressive tax system causing after-tax (disposable) income to be higher than it would otherwise be. Therefore aggregate demand falls less, making the recession less severe.

      The more progressive an income tax system, the greater the stabilising effect on economic activity.
    - ***Unemployment benefits***. In a recession, as real GDP falls and unemployment increases, unemployment benefits rise. If there were no unemployment benefits, unemployed workers' spending would fall significantly, putting a strong downward pressure on consumption spending and aggregate demand. The presence of unemployment benefits means that as workers become unemployed, their consumption will be maintained to some extent as their benefits partially replace their lost income, thus lessening the downward pressure on aggregate demand. In an expansion, unemployment benefits are reduced as unemployment falls; therefore, consumption increases less than it would in the absence of unemployment benefits.

      You should note that a progressive tax system and unemployment benefits cannot by themselves stabilise the economy and eliminate inflationary and recessionary gaps on their own. They can only help make economic fluctuations milder.

A note on monetary versus fiscal policy for demand management

Most economists believe that due to the advantages of monetary policy discussed earlier, this should play a more important role in demand management. The fact that it is incremental, flexible, and easily reversible, relatively free of political constraints, and with shorter time lags, makes it more appropriate as a tool to influence aggregate demand. By contrast, fiscal policy has an important role to play in deep recessions where even highly expansionary monetary policy may become ineffective (see for example Real world focus 13.1).

TEST YOUR UNDERSTANDING 13.7

1 Examine the strengths and weaknesses of fiscal policy and discuss which of these you think are more important.

2 (HL only)

  a Outline the meaning of automatic stabilisers.

  b Explain how

    i unemployment benefits, and

    ii a progressive tax system work to stabilise increases and decreases of aggregate demand.

3 a Use a diagram to explain crowding out.

  b Outline the disagreement over its importance when the economy is in recession.

4 Outline why decisions on whether or not to use fiscal policy measures depend on more than just economic considerations.

5 Explain why monetary policy is used more often than fiscal policy as a demand management tool.

REAL WORLD FOCUS 13.1

US fiscal policy and recession: two views

In February 2009, the Obama administration in the United States approved a fiscal stimulus package of $787 billion to support the US economy that was in recession. This had the effect of increasing the budget deficit. In spite of the fiscal measure, unemployment grew to over 10%. In the meantime, due to expansionary monetary policy, interest rates had fallen to nearly zero, suggesting that low interest rates were not working to increase aggregate demand.

Figure 13.5: Lakewood, Colorado, USA. Workers install solar panels funded by a federal (national) fiscal stimulus programme in 2010 to support the US economy

A Keynesian view on appropriate fiscal policy

The recession was deeper than originally estimated; therefore, a new fiscal stimulus package is required to supplement the first one. Recession involves low aggregate demand, and the government must step in by increasing its own spending. Through the multiplier (see Section 13.4), this works to increase spending in the economy by a multiplied amount, thus directly increasing *AD* and real GDP. Without the initial expansionary fiscal policy, the recession would have been deeper and longer, and unemployment much higher. The government's budget deficit increased, but this is unavoidable in recession, because of automatic stabilisers that work to lower taxes and increase government spending on unemployment benefits.

CONTINUED

A monetarist/new classical view on appropriate fiscal policy

The government should not have implemented the first fiscal stimulus package at all. It could have focused on reducing the budget deficit; this would improve business confidence, which would increase investment, and therefore economic growth. Otherwise, it could have implemented tax cuts, which are preferable to increased government spending, because deficit spending leads to higher interest rates that crowd out private investment spending. This reduces the expansionary effect of government spending. The increase of unemployment to over 10% suggests that the stimulus package did not work as expected.

A Keynesian response

The idea of reducing the budget deficit in a recession does not make sense. A household that spends more than its income can cut back on its spending, but if a government does that, the effect will be to reduce output and incomes, and increase unemployment, possibly making it even more difficult for the government to pay back its debts in the future. It is very unlikely that higher interest rates will crowd out investment in a recession when interest rates are already very low. Also, tax cuts have a smaller effect on increasing *AD*, because part of the tax cut is lost as increased saving; and if consumer confidence is very low, consumers may not spend their higher after-tax incomes and save most of it.

A monetarist/new classical response

Increased government spending adds to a large government that likely increases government intervention, which is inefficient, and acts as a drag on the economy. The focus should be on stimulating the economy without increasing the size of the government. Tax cuts offer this expansionary stimulus and at the same time create incentives for people to work more, thus possibly leading to lower unemployment without adding to government spending.

Applying your skills

1 Explain the meaning of
   a automatic stabilisers,
   b deficit spending, and
   c crowding out.
2 What assumptions about the macroeconomy are made by
   a Keynesian economists, and
   b monetarist/new classical economists, which lead them to differing conclusions about appropriate fiscal policy?

13.4 The Keynesian multiplier (HL only)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the Keynesian multiplier using the marginal propensity to consume, marginal propensity to save, marginal propensity to tax, marginal propensity to import (AO2)
- calculate the multiplier using (AO4) $\frac{1}{1-MPC}$ or $\frac{1}{MPS+MPT+MPM}$
- Using the Keynesian multiplier calculate the effect on GDP of a change in an injection (investment, government spending, exports) (AO4)

The nature and importance of the multiplier

Introducing the multiplier

Suppose there is an increase in one of the components of aggregate demand, due to a change in C, or I, or G, or $X-M$ expenditures.[3] This will produce an increase in aggregate demand, and an increase in real GDP. Yet the final increase in real GDP will most likely be greater than the initial increase in expenditures. The reason for this can be found in the **Keynesian multiplier**, defined as the change in real GDP divided by the initial change in expenditure:

$$multiplier = \frac{\text{change in real GDP}}{\text{initial change in expenditure}}$$

so that

$$\text{initial change in expenditure} \times \text{multiplier} = \text{change in real GDP}$$

As a rule, the multiplier >1; therefore, the change in real GDP is likely to be greater than the initial change in expenditure.

The multiplier is attributed to John Maynard Keynes, and is often referred to as the 'Keynesian multiplier'. It is important because it shows that whenever there is a change in a component of *AD*, there is likely to be a multiplied effect on real GDP. It is useful for policy-makers who often try to influence the level of aggregate demand in order to affect the level of real GDP and unemployment.

Understanding the multiplier in terms of leakages and injections

Why does a change in expenditure produce a larger change in total aggregate demand and real GDP? The explanation is that the initial change in expenditure produces a chain reaction of further expenditures, with the effect of increasing *AD* and real GDP to a value greater than the initial expenditure.

Assume an initial increase in investment spending of $8 million (due to business optimism or any other factor affecting investment spending). This spending of $8 million results in an increase in real GDP of $8 million. However, the story does not end there, because the $8 million increase in real GDP produces a further chain of spending, called *induced spending* (it is induced, or caused by the change in real GDP). The $8 million increase in investment spending is used by businesses to pay for materials, equipment, labour, etc., and all this spending translates into income for owners of the factors of production, who use it to increase their consumption spending. As consumption spending increases, it results in a further increase in real GDP and incomes, which produce more consumption spending. This process continues, increasing real GDP beyond the amount of the initial investment of $8 million.

To calculate the value of the multiplier, we must look at consumer spending more carefully. We know from the circular flow model that a portion of income flows out of the expenditure flow as leakages: consumers save part

[3] The change must be *autonomous*, meaning it has not been created by a change in income.

of their income, they pay taxes to the government, and they buy imported goods and services. The remaining part of income is spent on buying domestic goods and services, called consumption expenditures. This introduces us to a new concept, the **marginal propensity to consume**, abbreviated as *MPC*, defined as the fraction of additional income that households spend on consumption of domestically produced goods and services. For example, if the *MPC* is $\frac{3}{4}$, this means that given an increase in national income of \$10 million, $\frac{3}{4}$ of this, or \$7.5 million is consumption expenditure, and the remaining $\frac{1}{4}$ of income, or \$2.5 million, leaks out in the form of saving, taxes and spending on imports.

Corresponding to the marginal propensity to consume (*MPC*) is the **marginal propensity to save** (*MPS*, or fraction of additional income saved), the **marginal propensity to tax** (*MPT*, or fraction of additional income taxed), and the **marginal propensity to import** (*MPM*, or fraction of additional income spent on imported goods and services). Note that the $MPC + MPS + MPT + MPM = 1$. To see why, imagine that national income increases by \$1. It follows that the fractions of the \$1 that will be consumed, saved, spent on taxes and spent on imports will add up to \$1.

Assuming, as in Table 13.3 that the $MPC = \frac{3}{4}$, we can determine the value of the multiplier. The initial increase in investment spending of \$8 million results in an equivalent increase in income (or real GDP) of \$8 million. Since the $MPC = \frac{3}{4}$, this results in \$6 million of consumption expenditure ($\frac{3}{4} \times$ \$8 million = \$6 million). In the second round of income changes, the induced consumption expenditure of \$6 million leads to an equivalent increase in income of \$6 million, which when multiplied by the *MPC* produces new induced consumption spending of \$4.5 million. This process continues, with induced consumption spending and changes in income getting smaller and smaller until finally they drop to zero. Adding up all changes in income we arrive at a total increase of \$32 million. This is the amount by which real GDP has increased. This increase is equal to the initial change in investment spending of \$8 million plus the total increase in induced consumption spending of \$24 million.

Initial increase in investment expenditure of \$8 million:	Change in income (real GDP) (\$ million)	Induced change in consumption expenditure (\$ million)
1st round	8	$\frac{3}{4} \times 8 = 6$
2nd round	6	$\frac{3}{4} \times 6 = 4.5$
3rd round	4.5	$\frac{3}{4} \times 4.5 = 3.38$
4th round	3.38	$\frac{3}{4} \times 3.38 = 2.5$
(process continues an infinite number of times)		
Total	32	$\frac{3}{4} \times 32 = 24$

Table 13.3: Determining the value of the multiplier with the MPC = $\frac{3}{4}$

It thus follows that in this example, the value of the multiplier is:

$$\text{multiplier} = \frac{\text{change in real GDP}}{\text{initial change in expenditure}}$$

$$= \frac{\$32 \text{ billion}}{\$8 \text{ billion}} = 4$$

Alternatively, we can say that:

$4 \times \$8 \text{ million} = \$32 \text{ million}$

It is clear from the table that the value of the multiplier of 4 depends on the induced changes in consumption, which depend on the value of the *MPC*, assumed here to be $\frac{3}{4}$. The relationship between the multiplier and the *MPC* is:

$$\text{multiplier} = \frac{1}{1 - MPC}$$

If the $MPC = \frac{3}{4}$:

$$\text{multiplier} = \frac{1}{1 - \frac{3}{4}} = 4$$

Therefore, if we know the value of the MPC, we can calculate the value of the multiplier.

Now we know from the above that

$MPC + MPS + MPT + MPM = 1$

Rearranging this expression, we can write:

$1 - MPC = MPS + MPT + MPM$

We can therefore rewrite the multiplier as:

$$\text{multiplier} = \frac{1}{1 - MPC} = \frac{1}{MPS + MPT + MPM}$$

> The value of the multiplier is given by $\frac{1}{1-MPC}$, which is equivalent to:
>
> $\frac{1}{MPS + MPT + MPM}$.
>
> Therefore, if we know the value of the *MPC*, we can calculate the value of the multiplier. Alternatively, if we know the value of the *MPS*, *MPT* and *MPM*, we can calculate the value of the multiplier.

Based on these expressions, we can arrive at the following conclusions: *the larger the MPC, the smaller the value of the denominator of the first fraction, and so the greater is the multiplier*. Therefore, the greater the proportion of income spent on consumption, the greater the multiplier. Alternatively, we can see from the second fraction that *the smaller the leakages from the spending stream, the greater the multiplier*. Therefore the smaller the saving, or the level of taxes, or the volume of imports, the larger will be the size of the multiplier.

Whereas we calculated the multiplier based on an increase in investment spending, the same result would be obtained given an initial increase in any injection into the income flow, whether it is I, G, or X − M (as well as changes in C that have not been caused by changes in income).

Everything that has been said about the multiplier in relation to increases in expenditure applies equally to decreases in expenditures. Therefore, in the example above, if we had looked at a decrease in investment expenditure of \$8 million, and an MPC of $\frac{3}{4}$, there would result a decrease in GDP of \$32 million (= \$8 million × 4).

For example, Greece experienced an extremely deep recession in the period 2009 – 2018. This recession was much more serious than had been expected because the International Monetary Fund (IMF, see Chapter 20) underestimated the size of Greece's multiplier. Therefore in view of the actual size of the multiplier, spending cuts led to far greater decreases in real GDP than were anticipated (see Real World Focus 17.1).

Calculating the multiplier and its effects on real GDP

Suppose a country with a real GDP of £135 billion and an MPC of $\frac{4}{5}$ experiences an increase in exports of £2 billion. What is the change in real GDP, and the final value of real GDP?

To answer this question, we must first find the multiplier

$$\frac{1}{1-MPC} = \frac{1}{1-\frac{4}{5}} = \frac{1}{\frac{1}{5}} = 5$$

Therefore, there will be an increase in real GDP of £2 billion × multiplier = £2 billion × 5 = £10 billion.

Therefore, the final value of real GDP will be £135 billion + £10 billion = £145 billion.

If there had been a *decrease* in exports of £2 billion, there would result a £10 billion *decrease* in real GDP, thus making the final value of real GDP = £135 billion − £10 billion = £125 billion.

The multiplier, aggregate demand and real GDP

How the multiplier relates to aggregate demand

Using the example of an \$8 million increase in investment spending, we can see its effects on aggregate demand in Figure 13.6. The total aggregate demand shift is divided into two parts. The first part is the \$8 million increase in investment spending, called *autonomous spending*, meaning it has not been caused by a change in income. The second part is the effects on aggregate demand of the multiplier, which is \$24 million of *induced spending*, meaning spending caused by changes in income. The total effect on aggregate demand is the sum of autonomous plus induced spending, or \$32 million. This is equivalent to taking the initial change in autonomous investment spending and multiplying it by the multiplier: \$8 million × 4 = \$32 million.

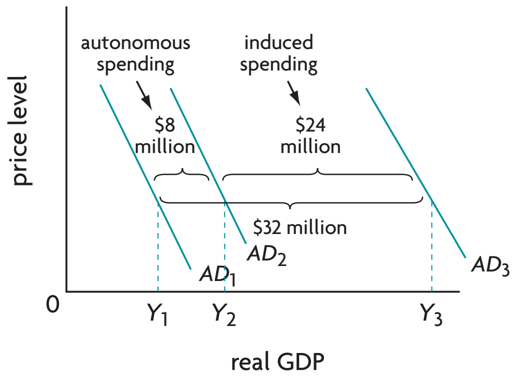


Figure 13.6: Aggregate demand, real GDP and the multiplier in the Keynesian model

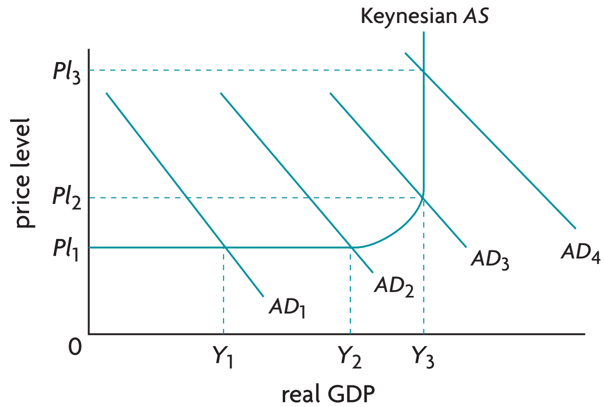


Figure 13.7: How the effect of the multiplier changes depending on the price level

All the factors listed in Table 9.1 (Chapter 9) under 'Shifts in the aggregate demand curve' can cause a change in spending resulting in a multiplier effect. All these factors involve changes in autonomous spending, because all are unrelated to income. (Remember that the factors listed in Table 9.1 are non-income factors.) This means that *the multiplier effect can only be initiated by a change in spending that is not caused by a change in income*.

We are now in a position to understand why this is so. Consider the *AD-AS* model in Figure 9.11. Each diagram shows an economy that is *in equilibrium*. This equilibrium determines a particular level of national income or real GDP. Since the economy is in equilibrium, *it is impossible for national income (or real GDP) to change unless something acts upon it from outside the system*. This 'something' must be unrelated to income, and can be any of the factors listed in Table 9.1, which are autonomous.

In our example, the outside change was autonomous investment spending of $8 million. This outside factor caused a change in income, and only then was it possible for the change in income to cause changes in consumption and aggregate demand; these are the induced changes shown in Figure 13.6 as the shift from AD_2 to AD_3.

The effect of the multiplier in relation to the price level

In order for the multiplier to have the greatest possible effect on real GDP, it is necessary that the price level is constant. We can see why in Figure 13.7, which shows the Keynesian *AD-AS* model with three equal *AD* curve shifts: from AD_1 to AD_2, then to AD_3, and finally to AD_4. The horizontal distance between each *AD* curve is identical. However, each shift occurs in a different section of the *AS* curve. The shift from AD_1 to AD_2 is in the horizontal part where the price level is constant, and *the increase in real GDP from Y_1 to Y_2 is exactly equal to the increase in aggregate demand*. Here we have the full multiplier effect. The shift from AD_2 to AD_3 occurs in the upward-sloping part of the *AS* curve, where the increase in real GDP, from Y_2 to Y_3, is smaller, because of the increase in the price level. The shift from AD_3 to AD_4 occurs in the vertical part of the *AS* curve, and results in no change at all in real GDP. The increasing price level has absorbed the entire multiplier effect, which here is zero.

In the monetarist/new classical model, increases in aggregate demand always (both in the short run and in the long run) lead to increases in the price level, therefore in this model it is never possible for real GDP to increase by the full amount of the increase in *AD*.

The multiplier is based on Keynesian thinking, which emphasises the point that in a recessionary gap (in the horizontal section of the *AS* curve), unemployed resources and spare capacity allow aggregate demand to increase without putting an upward pressure on the price level. Hence any autonomous increase in spending leads to a substantially larger increase in real GDP.

Therefore, when we use the multiplier to calculate the effects on real GDP of a change in autonomous spending, *we are presupposing a constant price level*.

TEST YOUR UNDERSTANDING 13.8 (HL ONLY)

1 a Define the multiplier.

b Outline why changes in real GDP are likely to be larger than the initial change in spending by a component of aggregate demand.

c Using a hypothetical numerical example explain why the multiplier is important.

2 a Define the marginal propensity to consume (*MPC*).

b Outline why the *MPC* is important in determining the size of the multiplier.

c Outline the role of leakages in determining the size of the multiplier.

3 Calculate the multiplier when the MPC is

a $\frac{4}{5}$

b $\frac{3}{4}$

c $\frac{2}{3}$

d $\frac{1}{2}$

4 Based on your answers to question 3, outline what we can conclude about the relationship between the size of the *MPC* and the size of the multiplier.

5 Calculate the value of the multiplier when

a the $MPS = MPT = MPM = 0.1$, and

b the $MPS = 0.13$, $MPT = 0.12$, and the $MPM = 0.15$.

6 Based on your answers to question 5, outline what we can conclude about the relationship between the size of $MPS + MPT + MPM$ and the size of the multiplier. Explain the reasoning behind this relationship.

7 In a country with a real GDP of $50 billion and an $MPC = \frac{2}{3}$, find the change in real GDP and the final value of real GDP (assuming a constant price level) for each of the following:

a an increase in net exports (exports minus imports) of $2 billion,

b a fall in investment spending of $3 billion,

c an increase in government spending of $7 billion, and

d a decrease in consumption spending of $1.5 billion.

8 Answer all the parts of question 7 assuming that the $MPS + MPT + MPM = \frac{1}{4}$.

9 **Optional** Using a diagram showing the Keynesian *AD-AS* model, show the effects of the multiplier when

a the price level is constant, and

b the price level is increasing.

13.5 Further topics on the multiplier and Keynesian economic theory (Supplementary material recommended for HL only)

If you are interested in exploring these issues further, you will find a discussion in the digital coursebook. The first part examines the relationship between the multiplier and fiscal policy more closely. The second part presents the Keynesian cross model which allows a better understanding of aggregate demand and its relationship to aggregate output (real GDP), and is also useful background to a deeper understanding of the multiplier.

13.6 Supply-side policies

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the goals of supply-side policies (AO2)
  - long-term growth through increases of productive capacity of the economy
  - improve competition and efficiency
  - labour market flexibility to reduce labour costs and unemployment
  - reduce inflation to improve international competitiveness
  - improve firm incentives to invest in innovation by decreasing costs
- explain market-based policies including (AO2)
  - policies to promote competition – deregulation, privatisation, trade liberalisation, anti-monopoly regulation
  - labour market policies – reduce power of labour unions, lower unemployment benefits, eliminate minimum wages
  - incentive related policies – cuts in personal income taxes, business taxes, capital gains taxes)
- draw a labour market diagram to show the minimum wage (AO4)
- explain interventionist policies including (AO2)
  - education, training
  - access to more and better health care
  - research and development
  - provision of infrastructure
  - industrial policies
- draw an *AD-AS* diagram with *LRAS* to show effects of supply-side policies (AO4)
- explain the demand-side effects of supply-side policies (AO2)
- explain the supply-side effects of demand-side policies (AO2)
- discuss constrains on supply-side policies (AO3)
  - market-based – effects on equity, time lags, vested interests, impact on the environment
  - interventionist – costs, time lags
- discuss strengths of supply-side policies (AO3)
  - market-based – improved resource allocation, no burden on the government budget
  - interventionist – direct support of sectors targeted to support growth
- evaluate supply-side policies with respect to promoting low unemployment, low and stable rate of inflation and growth (AO3); see Section 13.7

Goals of supply-side policies

The goals of supply-side policies include the following:

- **Promote long-term growth by increasing the productive capacity of the economy.** The main objective is to increase potential output, shown by a steeper long-term growth trend in the business cycle diagram, or rightward shifts of the *LRAS* curve (or Keynesian *AS* curve).
- **Improve competition and efficiency.** The objective is to make the economy more responsive to the market forces of demand and supply so as to increase efficiency in production.
- **Reduce costs of labour and reduce unemployment through greater labour market flexibility.** Greater labour market flexibility means making the labour market more responsive to the market forces of demand and supply so as to reduce unemployment as well as labour costs.
- **Increase incentives of firms to invest in innovation by lowering costs of production.** Higher after-tax profits through lower costs of production as well as lower taxes provide firms with incentives to engage in research and development that increases the productive capacity of the economy resulting in greater increases in productive capacity and growth in potential output.
- **Reduce inflation to improve international competitiveness.** Increases in potential output reduce inflationary pressures in the economy, thus making exports more competitive in global markets.

There are two types of supply-side policies: market-based and interventionist. Market-based policies emphasise the importance of well-functioning competitive markets, and are usually favoured by monetarist/new classical economists. Interventionist policies rely on government intervention to achieve growth in potential output, and are usually favoured by economists influenced by Keynesian thinking.

Market-based supply-side policies

In the early 1980s, some highly influential monetarist/new classical economists in the United Kingdom and the United States began to emphasise the view that growth in real GDP depends on the supply side of the economy. This view was adopted by the government headed by Margaret Thatcher in the United Kingdom, and by the government under Ronald Reagan in the United States. Since then, many governments throughout the world have pursued policies influenced by market-based supply-side thinking. In this view, the economy's real GDP tends automatically towards long-run full employment equilibrium and potential GDP (see Chapter 9). The focus of government policies should therefore be to create conditions that allow market forces to work well.

This perspective suggests that an economy pursuing supply-side policies will be able to achieve rapid growth, price stability and full employment all at the same time. As the economy tends towards full employment equilibrium, it automatically eliminates recessionary and inflationary gaps. If increases in aggregate supply match increases in aggregate demand so that the *LRAS* and *SRAS* curves[4] shift by the same amount as the *AD* curve, there need not be any price level increases. See Figure 11.3(a) in Chapter 11.

Market-based supply-side policies can be grouped under three headings:

1. Encouraging competition
2. Labour market reforms
3. Incentive-related policies.

Encouraging competition

Greater competition among firms forces them to reduce costs, contributing to greater efficiency in production and improving resource allocation, with the possible added benefit of improving the quality of goods and services. These benefits will allow potential output to increase and the *LRAS* curve to shift to the right.

- **Privatisation.** **Privatisation**, involving a transfer of ownership of a firm from the public to the private sector, can increase efficiency due to improved management and operation of the privatised firm. This is based on the argument that government enterprises are often inefficient due to bureaucratic procedures, high administrative costs and unproductive workers, because they do not face incentives to lower costs and maximise profits.
- **Deregulation.** **Deregulation** involves elimination or reduction of government regulation of private sector activities, based on the argument that government regulation stifles competition and increases

[4] See Chapter 9 Section *The relationship between the* SRAS *and* LRAS *curves in the monetarist/new classical model* for an explanation of why *SRAS* also shifts.

inefficiency. There are two main types of regulation (and deregulation): economic and social. 'Economic regulation' involves government control of prices, output, and other activities of firms, offering them protection *against competition*. In the last two to three decades, many countries have moved toward removal of government regulations, and hence economic deregulation. A main form of deregulation has been to allow new, private firms to enter into monopolistic or oligopolistic industries, thus forcing existing firms to face competition. The objective has been to increase efficiency, lower costs and improve quality. Industries affected include transport, airlines, television broadcasting, telecommunications, natural gas, electricity, financial services and others. 'Social regulation' involves protecting consumers against undesirable effects of private sector activities (many of these involve negative externalities) in numerous areas, including food, pharmaceutical and other product safety, worker protection against injuries, and pollution control. In contrast to economic regulation, social regulation is being strengthened in many countries in the interests of public safety. Some economists, however, argue that social regulation is excessive, giving rise to costly and inefficient bureaucratic procedures, paperwork and unnecessary government interference, and should therefore be reduced.

- **Contracting out to the private sector.** This is a policy option whereby governments make a contractual agreement with private firms to provide goods and services for the government (see Chapter 6). Examples include public goods, information technology, human resources management and accounting services. These result in increased competition as private firms compete with each other to get contracts with the government.
- **Anti-monopoly regulation.** Increased competition can result from restricting market power of firms by enforcing anti-monopoly legislation, by breaking up large firms that have been found to engage in monopolistic practices into smaller units that will behave more competitively, and by preventing mergers between firms that might result in too much market power. Greater scope for the forces of supply and demand may result in increased efficiency, lower costs and improved quality.
- **Trade liberalisation.** International trade between countries has become freer (*liberalised*) in recent decades due to reductions in trade barriers. Free or freer trade increases competition between firms both domestically and globally, which can result in greater efficiency in production and an improved allocation of resources (we will study this in Chapter 14).

Labour market reforms

Labour market reforms are sometimes referred to as increasing *labour market flexibility*, or reducing *labour market rigidities* discussed in Chapter 10 as one cause of structural unemployment. Labour market reforms are intended to get rid of rigidities by making labour markets more competitive, making wages respond to the forces of supply and demand, lowering labour costs and increasing employment by lowering the natural rate of unemployment. Lower costs of production can lead to increased profits, which in turn may result in greater investment by firms, increased R&D, increased capital goods production, and therefore increases in potential output (economic growth).

Labour market reforms include:

- **Abolishing minimum wage legislation.** Elimination or reduction of the legal minimum wage it is argued, reduces unemployment by allowing the equilibrium wage to fall. The benefits would include lower unemployment; greater firm profits, as wage costs would be lowered; more investment and economic growth. This can be seen in Figure 10.1(b) (Chapter 10).
- **Weakening the power of labour (trade) unions.** A **labour union**, or trade union, is an association of workers in a particular profession, whose objective is to improve working conditions and defend rights of workers, representing its members in negotiations with employers. Unionised labour frequently succeeds in securing high wage increases; if labour unions are weakened, wages will be more responsive to the forces of supply and demand, and will therefore be more likely to fall in if there is unemployment. This would also lead to the same benefits as abolishing minimum wage legislation.
- **Reducing unemployment benefits.** It is argued that unemployment benefits have the unintended effect of reducing the incentive to search for a new job, causing some unemployed workers to remain unemployed for longer periods than necessary. Therefore, reducing unemployment benefits is expected to lower unemployment, as it would encourage the unemployed to look for work. This could work to reduce the natural rate of unemployment.
- **Reducing job security.** Many countries have laws protecting workers against being fired, making it costly for firms to fire workers because of high

levels of compensation that must be paid to the worker being laid off. It is argued that reducing workers' job security by making it easier and less costly for firms to let go of workers has the effect of increasing employment, because firms are more likely to hire new workers if they know they can fire them easily and without cost if they are no longer needed. In addition, reducing job security would decrease firms' labour costs because of the lower costs of firing, and would therefore increase profits, investment and economic growth.

Incentive-related policies

Incentive-related policies involve cutting various types of taxes, which are expected to change the incentives faced by taxpayers, whether firms or consumers.

- **Lowering personal income taxes.** As we know, cuts in personal income taxes can increase aggregate demand. Supply-side economists argue that changes in personal income taxes have an even greater impact on aggregate supply because they *lead to higher after-tax incomes, creating an incentive for people to provide more work:* this can happen through an increase in the number of hours worked per week; an increase in the number of people interested in finding work (who were formerly not interested in working); an increase in the number of years worked, as people may decide to retire later; a decrease in unemployment as unemployed workers choose to shorten the duration of their unemployment. All these factors may work to shift the *LRAS* curve to the right, increasing potential output.
- **Lowering taxes on capital gains and interest income.** Taxes on **capital gains** are taxes on profits from financial investments (such as stocks and bonds) or from buying and selling real estate. If the taxes on capital gains and on income from interest on savings deposits are reduced, people may be more motivated to save, thus increasing the amount of savings available for investment. More investment means a greater production of capital goods and an increase in potential output.
- **Lowering business taxes.** Lower taxes on business profits (corporation taxes) can work to increase aggregate demand by increasing investment spending. Supply-side economists argue that cutting taxes on firms' profits is a supply-side measure because increases in the level of after-tax profits mean that firms have greater financial resources for investment and for pursuing technological innovations through more R&D, resulting in greater potential output.

Interventionist supply-side policies

Interventionist supply-side policies presuppose that the free market economy alone cannot achieve the desired results in terms of increasing potential output, and therefore government intervention is required.

Investment in human capital: education and health services

Investment in human capital can take the following important forms:

- **Training and education.** More and better training and education lead to an improvement in the quality of labour resources, increasing the productivity of labour, which is one of the key causes of economic growth (see Chapter 11). Education also has numerous positive externalities, justifying government intervention. Public training and education programmes can assist workers to become more employable, thus reducing the natural rate of unemployment. Specific measures include setting up retraining programmes for structurally unemployed workers to obtain skills in greater demand; assisting young people to pursue training and education through grants or low interest loans; direct government hiring and provision of on-the-job training; providing grants to firms that offer on-the-job training; offering subsidies to firms that hire structurally unemployed workers; assisting workers to relocate to geographical areas where there is a greater demand for labour through grants and subsidies (such as provision of low-cost housing); providing information on job availability in various geographical areas; establishing government projects in the depressed areas that result in new employment creation.
- **Improved health care services and access to these.** When workers (and the general population) have access to good quality health care services, they become healthier and more productive. More and improved health care services and access to these by the working population is another factor leading to improvements in the quality of labour resources, increasing the economy's potential output. Health care also has many positive externalities, justifying government intervention.

Investments in human capital result in an increase in aggregate demand over the short term, and over the longer-term lead to increases in potential output, by shifting the *LRAS* or Keynesian *AS* curves to the right.

Investment in new technology: research and development

Research and development (R&D) is the fundamental activity behind the development of new technologies, resulting in new or improved capital goods (physical capital), which is another important cause of increases in potential output and economic growth (see Chapter 11). R&D also has positive production externalities, thereby justifying government intervention.

Governments in many countries around the world are therefore heavily involved in R&D. In addition, governments often provide incentives to private sector firms to engage in R&D activities; these usually take the form of tax incentives, as well as the granting of patents for the protection of inventions.

Government spending in support of new technology development leads to increases in aggregate demand over the short term and increases in potential output over the longer term shifting the *LRAS* or Keynesian *AS* curves to the right.

Investment in infrastructure

Infrastructure is a type of physical capital, and therefore results from investment; it includes power, telecommunications, roads, dams, urban transport, ports airports, irrigation systems, etc. Many types of infrastructure qualify as merit goods or public goods, thereby justifying government intervention. More and better infrastructure increases efficiencies in production as it lowers costs. Good roads, railway and other transport systems, for example, save time and effort spent in transporting goods and services, allowing more output to be transported and costs to be lowered. The availability of effective telecommunications permits faster and easier communications, enabling economic activities to be carried out more efficiently. More and better infrastructure improves labour productivity. Investments in infrastructure therefore work to increase aggregate demand over the short term, but they also contribute to increases in potential output and *AS* increases over the longer term.

Industrial policies

Industrial policies are government policies designed to support the growth of the industrial sector of an economy. They include:

- **Support for small and medium-sized enterprises or firms (SMEs).** This may take the form of tax exemptions, grants, low-interest loans and business guidance. These measures provide support for the private sector, promoting efficiency, more capital formation, more employment possibilities and therefore increases in aggregate demand as well as potential output.
- **Support for 'infant industries'.** 'Infant industries' are newly emerging industries in developing countries, which sometimes receive government support in the form of grants, subsidies, tax exemptions, and tariffs or other forms of protection against exports (see Chapter 14). This also provides support for growth of the private sector and increases in aggregate demand and growth in potential output.

TEST YOUR UNDERSTANDING 13.9

1 a Explain the goals of supply-side policies.

  b Outline the two categories of supply-side policies noting how they differ in their general focus.

2 Using an appropriate diagram based on the *AD-AS* model, illustrate and explain the expected impacts of supply-side policies on real GDP, the price level and unemployment. (You may use the Keynesian or monetarist/new classical model to illustrate.)

3 a Provide some examples of interventionist supply-side policies.

  b Using a diagram, explain how these policies affect aggregate demand over the short term and also increase *LRAS*.

  c Show these effects using the Keynesian *AD-AS* model.

4 Explain why supporters of market-based supply-side policies argue that by focusing on the supply side of the economy, it is possible to address the policy goals of economic growth, price stability and unemployment all at the same time.

5 Outline what advantages supply-side policies might have over demand-side policies in the event that an economy is experiencing stagflation (simultaneous inflation and unemployment with recession).

CONTINUED

6 Provide some examples of supply-side policies that aim to achieve each of the following objectives:

a increase competition,

b improve incentives, and

c make the labour market more responsive to supply and demand.

Overlaps between demand-side and supply-side policies

Demand-side policies are focused on influencing the short-term fluctuations of aggregate demand while supply-side policies are intended to affect long run aggregate supply. Yet there are some important overlaps between the two.

Demand-side effects of supply-side policies

Interventionist supply-side policies

When the government invests in education or training, health care services, research development, provision of infrastructure it increases the quantity or improves the quality of physical capital and human capital thus having an impact on *LRAS* (or Keynesian *AS*). However, these activities involve an increase in government spending, which has the effect of shifting aggregate demand to the right. These are interventionist supply-side policies with effects on the demand side of the economy.

Market-based supply-side policies

Incentive-related policies, including lower taxes on business profits, are intended to encourage firms to invest more in R&D, new technologies and new capital goods with the intention to increase *LRAS*. But any increase in investment leads to an increase in aggregate demand. Similarly, cuts in personal income taxes create incentives for workers to work more, thus increasing *LRAS*. Yet cuts in income taxes also increase disposable income which increase consumption, thus increasing aggregate demand.

Supply-side effects of demand-side policies

Fiscal policy focuses mainly on short-term demand management. However, it can also contribute to long-term growth of potential GDP. It can do so indirectly, by providing a stable macroeconomic environment (discussed under fiscal policy), and directly, by leading to aggregate expenditures that result in growth of potential GDP.

Direct supply-side effects of demand-side policies on potential output and long-term economic growth

Fiscal policy

- Government spending for provision of physical capital goods, such as infrastructure (roads and transport systems, telecommunications, harbours, airports, etc.), as well as on R&D, which improves technology and therefore the quality of capital goods, and increases the productivity of labour.
- Government spending for the development of human capital, such as training and education programmes as well as health care that improve the quality of the labour force and increase the productivity of labour.
- Provision of incentives to encourage investment by firms through lower business taxes thereby contributing to new capital formation and R&D that promotes technological innovations.

Monetary policy

While fiscal policy has the stronger ability to affect potential output, monetary policy can also affect it. A fall in interest rates encourages more spending by firms on capital goods, which increases their quantity. An increase in the quantity of any factor of production affects potential output, therefore a fall in interest rates has an impact on long term growth.

All these factors work to increase aggregate demand as well as potential output, thus supporting long-term economic growth. These effects can be seen in Figure 11.3 (Chapter 11). Suppose an economy is initially in equilibrium producing real output Y_1, and the government pursues a variety of demand-side fiscal policies, including increases in government expenditures on infrastructure, R&D and training and education, thereby increasing the quantity of capital goods, and improving the level of technology and the quality of labour force. These policies produce increases in aggregate demand over the short term, so that *AD* shifts from AD_1 to AD_2. However, these policies also impact on aggregate supply, because of the increase in the quantity of capital goods, the improvements in the quality of labour, etc., so that the *LRAS* and *SRAS* curves (part (a)) and the Keynesian *AS* curve (part (b)) also shift to the right increasing potential output over the long term.

Some demand-side policies have not only demand-side but also supply-side effects, and can affect long-term economic growth by increasing potential output. Their contribution to economic growth includes creating a stable economic environment, as well private investment spending and government spending, in turn leading to increases in potential output through new capital formation, increased R&D and technological improvements, and improvements in the quality of the labour force.

Evaluating supply-side policies

We turn to an evaluation of the effectiveness of supply-side policies.

Constraints on market-based supply-side policies

Time lags

The policies work after significant time lags, making their effects on *the supply side* of the economy (aggregate supply) over the longer term. This is because the activities set into motion (increased competition, labour market reforms, changing incentives) need time to materialise and affect potential output.

Possible unfavourable impact on unemployment

Market-based policies that focus on encouraging competition may well *increase* unemployment, at least over the short-term. In the case of privatisation, as privatised firms try to make their operations more efficient, they often try to cut costs by firing workers. Contracting out to the private sector leads to government job losses, and job losses for the country as a whole if projects are contracted out to firms in other lower cost countries. In addition, economic deregulation has frequently led to increased unemployment, due to increased competitive pressures that cause firms to fire workers in order to lower their costs. It is possible that increased unemployment on account of these policies may be short-term, and may be reversed over the longer term as the economy begins to benefit from the broader effects of supply-side policies.

Possible negative effects on equity

Market-based policies often have negative effects on equity. Greater competition may have a negative effect if it results in some unemployment, involving a loss of income. Labour market reforms include changes in legislation and institutions that provide protection for workers with very low incomes and with income uncertainties (minimum wage legislation, protection against being fired, unemployment benefits). Reduced protection results in lower incomes for some workers and increased job insecurity, and contributes to rising income inequalities. Minimum wage legislation, for example, is intended to protect unskilled workers on very low incomes. Unemployment benefits and job security are especially important to people on very low incomes who have nothing to fall back on if they are left unemployed.

In the case of incentive-related policies, tax cuts intended to create incentives to work, save and invest may also worsen income distribution. The argument that high taxes create disincentives to work and save applies mainly to higher income groups who face higher average tax rates; therefore, to reverse this problem, tax cuts must be designed to affect the after-tax incomes of higher income groups. Yet this would make the tax system less progressive, reducing the redistributive effects of personal income taxes and making income distribution less equal. In addition, since it is the wealthy who enjoy capital gains and earn most of the interest income and business profits, tax cuts in these areas will affect wealthy people by increasing their after-tax incomes more than they will affect lower income groups of the population.

Another point concerns the prices of products sold by privatised firms. If private firms have a degree of market power, they are likely to raise their prices over what the government used to charge (as well as restrict the quantity of output produced). As a result their products become less affordable, with damaging consequences for lower income groups, particularly if the privatised firms provide necessities or merit goods, such as utilities, including power, water supplies, sewage systems, etc. (This is a problem in some developing countries, where privatisation of services has made these unaffordable for very poor people.)

Negative impact on the government budget

Market-based policies do not need government funds to be implemented as they are based on private initiative. However, incentive-related policies in the form of tax cuts negatively affect the budget as they reduce tax revenues. They can therefore lead to a budget deficit or increase an existing one. (See Real world focus 13.2.)

Possible interference of vested interests

Vested interests are strong personal interests in something. Market-based policies often affect particular stakeholders in ways which are not in their best interests, and these groups therefore oppose and may prevent the policies from being implemented. For example, strong labour unions oppose any policy that would weaken their power, or would reduce or abolish the minimum wage. Workers in public (government-owned) enterprises may be opposed to privatisation as they fear for their jobs which may be cut when the privatised firm tries to reduce costs. Firms with strong market power in industries where there is little competition are likely to be strongly opposed to policies that promote competition by breaking up large firms with too much market power, or preventing mergers between firms that would like to have more market power.

Possible negative effects on the environment

Market-based policies focused on increasing competition (privatisation, deregulation) may have negative effects on the environment because of the increased scope for activities leading to negative externalities affecting the environment.

Constraints on interventionist supply-side policies

Time lags

Here too the policies work after significant time lags because of the time needed for investments, new human and physical capital, R&D, and so on to be realised and to take effect.

Negative impact on the government budget

Interventionist policies have negative effects on the government budget because they are heavily based on government spending. They can therefore create a budget deficit (or can increase the size of the deficit if there was one to begin with).

Strengths of market-based supply-side policies

Improved resource allocation

Market-based supply-side policies focus on improving the workings of the market system based on the operation of demand and supply, and these are expected to result in improved efficiency in resource allocation.

May not burden the government budget

Most market-based policies do not need government funds to be implemented as they are based on private initiative. (Tax cuts are an exception; see above.)

Ability to create employment

Market-based policies involving labour market reforms may also contribute to reducing the natural rate of unemployment by focusing on making the labour market more responsive to supply and demand (lower wages and production costs, easier hiring and firing, etc.).

Ability to reduce inflationary pressure

By increasing potential output supply-side policies are likely to reduce inflationary pressures over the longer term. As potential output increases with *LRAS* shifting to the right so as to match increases in aggregate demand, there will be little or no upward pressure on the price level. This can be seen in Figure 11.3 (Chapter 11).

The ability of market-based policies to reduce inflationary pressures can also be understood in terms of the focus on keeping firms' costs of production down through increases in efficiency (due to increased competition) and lower wage costs (due to increased labour market flexibility).

Strengths of interventionist supply-side policies

Direct support of sectors important for growth

The government selects particular sectors or activities to promote, which may be important for growth. For example if a country needs infrastructure such as road systems, ports or airports, government support for these can be crucially important for the effective functioning and growth of the economy.

Ability to create employment

Interventionist policies involving investments in education and training can make a direct impact on a reduction of unemployment by:

- enabling workers to acquire the skills, training and retraining necessary to meet the needs of employers (structural unemployment)
- providing assistance to workers to relocate (structural unemployment)
- providing information that reduces unemployment when workers are between jobs (frictional unemployment) or between seasons (seasonal unemployment).

Potential ability to reduce inflationary pressure

Interventionist policies may also reduce inflationary pressures by increasing potential output, as Figure 11.3 shows.

Possible positive effects on equity

Interventionist policies that focus on investments in human capital that are broadly distributed throughout the population are likely to have positive effects on equity over the longer term. The reason is that educated, skilled and healthy workers are more likely to be employed and be an active and productive part of society, with the result that income is likely to be relatively more equally distributed.

REAL WORLD FOCUS 13.2

The US reduces corporate taxes

In 2017, the United States enacted the largest ever reduction in corporate tax rates in US history, from 35% to 21%, bringing them to their lowest level in 50 years. This was justified on the grounds that increased investment due to the tax cut would result in more growth that would in turn lead to higher real incomes. According to the US President's economic advisor, 'With supply-side tax cuts . . . we are producing very significant growth with virtually no inflation.'[5]

Figure 13.8: Businessman about to cut a tax paper

Yet according to a major survey of corporations, the 'tax cut package appeared to have no major impact on businesses' capital investment or hiring plans'.[6] The International Monetary Fund (IMF) similarly concludes that the tax cuts have not brought forth the expected levels of investment.

In the meantime, corporate tax revenues fell by double the expected amount in 2018. Economists are debating the possible causes of this but have yet to identify them. The significant drops in tax revenues are enlarging the budget deficit.

The benefits of the tax cut went to corporate shareholders in the form of higher profits (dividends on their shares of stocks) that are not being reinvested. Yet 35% of corporate ownership is foreign. This means that 35% of the benefits of the tax cut is transferred abroad to investors outside the United States. Paul Krugman, US Nobel prize winning economist, notes that this amounts to more than $40 billion each year that is given away to wealthy foreign investors. He also notes that this amount is much higher than what the United States spends on foreign aid.

Sources: *www.bbc.com; www.latimes.com; www.cnbc.com; www.nytimes.com*

Applying your skills

1. Using a diagram, explain what the US President's economic advisor meant in the statement quoted in the first paragraph.
2. Referring to the topic of income inequality of Chapter 12, explain the possible connection between the cuts in corporate income taxes with increasing wealth inequality
3. The corporate tax cut has been hotly debated in the United States. Research this topic and present both sides of the debate.

[5] https://www.cnbc.com/2019/01/28/1point5-trillion-us-tax-cut-has-no-major-impact-on-business-spending-plans-survey.html

[6] https://www.cnbc.com/2019/01/28/1point5-trillion-us-tax-cut-has-no-major-impact-on-business-spending-plans-survey.html

Some ongoing debates

Most economists would agree that it is unlikely that any policy can yield positive results without some negative consequences. Most believe that interventionist and market-based policies should complement each other, and the particular mix of policies that should be used will likely be different according to each country's particular economic and social conditions (see for example Real world focus 13.3). However, major disagreements persist over the suitability and particular mix of policies for achieving long term economic growth. Here is a small sample of the debates.

The debate over increases in potential output

Economists generally agree that supply-side policies play an important role in increasing potential output. However, they disagree on whether interventionist or market-based polices are more effective. Supporters of interventionist policies argue in terms of the major advantages of targeted government support in areas such as investment, R&D, training and education, provision of credit on favourable terms (low interest rates, long repayment periods), and so on; they argue that the market is unlikely to provide them as needed. Moreover, industrial policies allow the government to support particular industries that are held to offer the greatest possibilities for growth in the future. They point to the experiences of a group of Asian countries (the 'Asian Tigers'; see Chapter 20) which achieved very high rates of growth by use of highly interventionist policies focusing on investments in human capital and industrial policies. They also point to the questionable growth performance of many developing countries that adopted market-based supply-side policies in the 1990s and beyond (see Chapter 20).

Supporters of market-based policies argue that government interference in the market may lead to inefficiencies and resource misallocation, whereas reliance on the market can achieve long-term growth while avoiding these disadvantages. A major argument against government intervention and industrial policies involves the idea that government interference may result in less efficient outcomes because of the influence of political pressures, lack of necessary information and unintended and unwanted consequences of government actions. Governments may lack the ability to choose the right industries to support, and incorrect choices will lead to a poor allocation of resources.

Supporters of market-based policies also note that interventionist policies rely heavily on government spending, using resources that might have better alternative uses elsewhere (opportunity costs). Governments require substantial amounts of tax revenues to be able to provide the support services, which means high taxes and a large government sector. High taxes act as disincentives to work, and a large government sector promotes inefficiencies.

The debate over incentive-related policies

Tax cuts (incentive-related policies) are among the more controversial market-based policies, because of their questionable effects on work, saving and growth of potential output. Tax cuts have both demand-side and supply-side effects. Some economists question the strength of the supply-side effects, believing these to be small compared to the impact on aggregate demand. Increases in disposable income due to cuts in personal income taxes may result in the decision to work less if people prefer to use their extra (after-tax) income to increase their time for leisure. Also, workers may decide to use their higher after-tax income to consume more rather than save, in which case the tax cuts may not significantly affect saving and investment. In the United States, for example, whereas there have been a series of tax cuts, savings are at their lowest point in the past century. In countries where tax cuts were implemented as supply-side policies (such as in the United Kingdom and the United States), economists disagree on whether or not these have worked to increase potential output. The reason is that whatever growth has occurred has been the result of both demand-side and supply-side effects of demand-side and supply-side policies, and it is very difficult to detect which particular policy has been responsible for each particular effect.

The debate over minimum wages

Based on the analysis of Figure 10.1(b) (Chapter 10), that shows unemployment resulting from a minimum wage, supporters of market-based policies argue in favour of reducing the minimum wage to reduce unemployment. However, there is some question over whether reducing minimum wages will produce this result. Many economists argue that paying workers a higher than equilibrium wage encourages them to work harder, increasing their productivity (the output produced per worker). Increased labour productivity causes firms to increase their demand for labour, which has the impact of increasing employment and justifying the higher wages. If this argument is correct, there would be little benefit for firms if governments cut the minimum wage. There are many studies that empirically support this argument.[7] According to some views it may be the case that the minimum wage may begin to have an impact on unemployment only when it approaches very high levels.

TEST YOUR UNDERSTANDING 13.10

1 Using examples, explain
   a some supply-side effects of demand-side policies, and
   b demand-side effects of supply-side policies.

2 Discuss advantages and disadvantages of interventionist supply-side policies, including
   a investment in human capital,
   b investment in new technology,
   c investment in infrastructure, and
   d industrial policies.

3 Discuss advantages and disadvantages of market-based supply-side policies, including
   a policies to encourage competition,
   b labour market reforms, and
   c incentive-related policies.

4 Explain why minimum wages may not lead to unemployment, as illustrated in a labour market diagram.

REAL WORLD FOCUS 13.3

'Flexicurity' in Denmark

The people of Denmark appear to have achieved a combination of things that many economists would consider impossible. The Danish economy has been one of the best performing in the world; its GNI *per capita* in ($PPP) ranks second in the EU; the poverty rate is among the lowest in the world; it has one of the most equal distributions of income in the world; and according to the 'World Map of Happiness', the Danes are the second happiest people in the world; all this while also paying among the highest taxes in the world, with one of the lowest tax evasion rates in the world.

Figure 13.9: Copenhagen, Denmark. Nyhavn New Harbour canal

Government officials from Europe and the United States have been travelling to Denmark to discover the secret of its success, which lies partly in the unique practice of 'flexicurity', derived from 'flexibility' and 'security'.

The flexibility part of 'flexicurity' is based on Denmark's highly flexible labour market. Workers can be easily fired with little prior notice, meaning they can also be easily hired (because there is little or no cost involved in firing). There is a very large turnover in the labour market, with 30% of the labour force switching jobs each year. Most of these switches are not due to lay-offs, but moving on to better jobs. Though this gives rise to some unemployment, Denmark has achieved one of the lowest unemployment rates in the European Union.

[7] See for example https://www.bloomberg.com/opinion/articles/2019-01-24/u-s-economy-higher-minimum-wages-haven-t-increased-unemployment
https://www.businessforafairminimumwage.org/news/00135/research-shows-minimum-wage-increases-do-not-cause-job-loss

CONTINUED

The security part of 'flexicurity' is based on Denmark's extensive social protection system. Once a worker is fired, they are is entitled to very generous unemployment benefits, amounting to 90% of the wage for a maximum of three years over a lifetime of work. This provides workers with the incentive to find a new job soon after they have been laid off.

One condition of receiving unemployment benefits is that workers must be available to take on a job that is offered to them through government job centres after 12 months of unemployment. In addition, the government provides free education and training to unemployed workers to help them easily find new jobs. Most workers belong to labour unions that work very closely with businesses to discover what skills and education employers require. This helps reduce the level of structural unemployment in the economy. Denmark has a highly skilled and educated labour force.

The security part of 'flexicurity' is also based on public provision of free education from kindergarten through university, free health care and hospitals, generous retirement pensions, housing subsidies for low-income earners and numerous other social benefits.

An additional possible explanation for Denmark's economic success is its market-oriented economy, based on free trade, competition, and limited government ownership or intervention in business. It is also considered to be the country with the least amount of bureaucracy and the shortest amount of start-up time for new firms in the European Union.

Denmark also has among the highest income tax rates in the world, with incomes taxed at nearly 56% on average. Personal income taxes are strongly progressive, and this contributes to the high degree of income equality. Also, it has a very high value added tax (VAT, an indirect tax) at 25%. High taxes are necessary to pay for the very generous unemployment benefits, free education and health care, and other merit goods provided by the government. On the other hand, business taxes are comparable with most other European countries.

Source: *https://economic-research.bnpparibas.com*

Applying your skills

1. Identify what policies in Denmark are useful for maintaining low rates of
   a. frictional unemployment, and
   b. structural unemployment.
2. Describe how a highly progressive tax system contributes to greater equality in income distribution (see Chapter 12).
3. It is often argued that highly flexible labour markets lead to greater income inequality. Identify the policies that Denmark uses to ensure this does not occur.
4. Denmark has combined high economic growth rates together with taxes that are among the highest in the world. Describe what this suggests about possible disincentive effects of very high taxes.
5. Denmark's unemployment benefits are among the most generous in the world, yet it has very low unemployment rates. Describe what policies are used to avoid possible disincentive effects of unemployment benefits toward work.
6. Identify the kind of unemployment involved by workers 'moving on to better jobs'.
7. Denmark has a unique mixture of interventionist and market-oriented supply-side policies that appear to contribute to its success. Explain what these are.

13.7 Evaluation of demand-side and supply-side policies to promote low unemployment, low and stable rate of inflation and economic growth

LEARNING OBJECTIVES

After studying this section you will be able to:

- evaluate monetary policy, fiscal policy and supply-side policies with respect to promoting low unemployment, low and stable rate of inflation and growth (AO3)

Policies for low unemployment

Different types of unemployment (see Chapter 10) require different kinds of policies for their solution. The main distinction is between cyclical (demand-deficient) unemployment and natural unemployment.

Cyclical unemployment

Since cyclical unemployment is caused by low or falling aggregate demand, measures to correct it involve expansionary demand-side policies, or monetary and fiscal policies. The intended effects of such policies are shown in Figure 13.2(a) and (b), where the economy is initially in a recessionary gap producing output Y_{rec}. Efforts by the government or central bank to shift AD from AD_1 to AD_2 are intended to increase real GDP Y_p representing potential output. As AD shifts to the right, the recessionary gap shrinks, and cyclical unemployment falls until it is eliminated at Y_p.

Monetary policy has the advantages that it can be used incrementally, and it can easily be reversed or changed in the event that aggregate demand increases too little or too much in response to lower interest rates. In addition, it is not subject to political constraints and does not have as long time lags as fiscal policy. It also does not lead to crowding out, and does not lead to increased government spending with larger budget deficits and government debt.

However, if the recession is deep, with low business and consumer confidence and fears of bankruptcies in the banking sector, lower interest rates may not lead to the needed increase in aggregate demand. Moreover, once interest rates reach zero they cannot be reduced further.

Fiscal policy has the ability to pull an economy out of deep recession with a high rate of cyclical unemployment. In a deep recession expansionary fiscal policy can complement monetary policy. Increased government spending, in particular, has a direct impact on aggregate demand. Government spending has the further advantage that it can be targeted to particular economic activities, such as spending on infrastructure, that have supply-side effects through their impact on potential output. Moreover, automatic stabilisers in the form of progressive income taxes and/or unemployment benefits make the recession less severe, and therefore make cyclical unemployment not as high as it would have been if these stabilisers were not present.

On the other hand, fiscal policy faces a number of disadvantages. It cannot fine tune the economy and is subject to major time lags, so that by the time the policy takes effect it may no longer be appropriate. In addition, since in a recession tax revenues fall due to rising unemployment and government spending increases on unemployment benefits, the government is likely to have a budget deficit resulting in increased government debt. Increased government borrowing may lead to higher interest rates possibly crowding out private investment.

Supply-side policies mainly affect potential output, whereas cyclical unemployment is due to insufficient aggregate demand. Market-based supply-side policies will not help reduce cyclical unemployment. However interventionist supply-side policies could be effective since they have demand-side effects as discussed earlier.

Natural unemployment

Structural unemployment is the most serious part of natural unemployment, and most economic policies intended to lower the natural rate of unemployment focus on this.

Demand-side policies are generally not appropriate. To see why, suppose that an economy is producing at the level of potential output, with unemployment equal to the natural rate. If aggregate demand is increased through fiscal or monetary policy, the natural rate of unemployment will fall temporarily; however, this will cause inflation (you can see this by using either the monetarist/new classical or Keynesian models). Policy-makers would therefore reduce aggregate demand to lower the rate of inflation and unemployment will fall once again to its natural rate. Alternatively, in the event that there is downward price and wage flexibility, the monetarist/new classical model shows that in the long run the economy will revert to long-run equilibrium with unemployment once more equal to the natural rate.

Fiscal policy however may have effects on natural unemployment because of its *supply-side effects*. These kinds of fiscal policy measures are included within interventionist supply-side policies.

Interventionist supply-side measures to reduce structural unemployment include setting up retraining programmes; support for re-training through grants and low interest loans; direct government hiring and provision of on-the-job training; grants to firms offering on-the-job training; subsidies to firms hiring structurally unemployed workers; grants and subsidies to assist relocation; information on job availability in various geographical areas; government projects in the depressed areas for employment creation.

Measures to reduce frictional unemployment aim at improving information flows between employers and job seekers, reducing the time a worker spends searching for a job. Improved information can result from the establishment of job centres, employment agencies and other methods of facilitating information exchanges, such as job websites.

Measures to reduce seasonal unemployment include provision of information to workers on jobs available during off-peak seasons in other industries.

The advantages of such policies are that they have a direct positive impact on reducing unemployment, without contributing to increased income inequalities and loss of job security. Disadvantages include the negative impacts on the government budget and opportunity costs of government spending.

Market-based supply-side measures include labour market reforms that increase labour market flexibility. As we know, reducing the minimum wage could potentially reduce unemployment by lowering wages of unskilled workers; weaker labour unions reduce the upward pressure on wages making it easier for firms to hire because of lower costs; reducing job security makes it easier for firms to hire because they can more easily fire; and reduction of unemployment benefits increase workers' incentives to find work.

These measures are aimed at structural, frictional and seasonal unemployment. The advantages of these policies are that they can reduce the natural rate of unemployment without negative effects on the government budget. The major disadvantages are that they contribute to income inequality and loss of protection for low-income workers.

Policies for a low and stable rate of inflation

It is important to bear in mind the distinction between demand-pull and cost-push inflation (see Chapter 10) as this determines the policies appropriate to deal with each one.

Demand-pull inflation

Since demand-pull inflation is caused by increases in aggregate demand, appropriate policies are contractionary demand-side policies, or monetary and fiscal policies that attempt to bring about a decrease in aggregate demand, so that AD_2 shifts toward AD_1 in Figure 13.3(a) and (b) bringing the economy back to potential output Y_p.

Monetary policy in the first instance is more appropriate as it can be used incrementally, and can easily be reversed or changed in the event that aggregate demand decreases too little or too much in response to higher interest rates. In addition it has the important advantage of shorter time lags compared with fiscal policy. The further advantage of no political constraints is more important here than in the case of dealing with unemployment, because the tax increases or government spending cuts called for by contractionary fiscal policy are politically highly unpopular, making it difficult for governments to undertake these policies.

On the negative side there may be conflict among government objectives. Inflation requires higher interest rates, but this may increase the exchange rate (appreciation), which will make imports cheaper and exports more expensive to foreigners. If the country has a trade deficit (more imports than exports), a currency appreciation may work to increase the size of the trade deficit, which is not desirable (see Chapter 16).

Fiscal policy may be useful to deal with rapid and escalating inflation as a complementary policy to monetary policy. Cuts in government spending could also have a direct impact on reducing aggregate demand. Finally, automatic stabilisers are expected to play a positive role as progressive income taxes and unemployment benefits work to make the inflation less severe.

However, cuts in government spending are highly unpopular, as are also increases in taxes. Fiscal policy is a cumbersome tool not well suited to fine tuning the economy. In addition time lags could make the policy inappropriate by the time it takes effect.

Supply-side policies cannot be used to deal with demand-pull inflation over short periods of time, because demand-pull inflation has causes lying on the demand-side, and supply-side policies work with a long time lag. However, over long periods, supply-side policies, whether interventionist or market-based do have the tendency to reduce inflationary pressures that might have demand-side causes, because they shift the *LRAS* or Keynesian *AS* curves to the right (see Figure 11.3) As long as the productive capacity of the economy is growing at least as fast as aggregate demand, inflationary pressures are kept under control.

Cost-push inflation

Cost-push inflation is caused by an increase in costs of production or supply-side shocks, causing a leftward shift in the *SRAS* curve, resulting not only in a higher price level but also a fall in real output and a rise in unemployment.

Demand-side policies are problematic, because, whereas the problem of inflation requires a decrease in aggregate demand, the problem of unemployment requires an increase in aggregate demand. In spite of this conflict central banks committed to a low rate of inflation use contractionary monetary policy (raising interest rates) to lower aggregate demand. This comes at the cost of more recession and therefore increased unemployment. Fiscal policy is not appropriate and is not generally used.

Supply-side policies may be used though there are no general solutions to the problem of cost-push inflation. Policies that can be pursued depend very much on the specific cause of the increase in costs. For example, if cost-push inflation is due to increases in wages, the appropriate solution may lie in supply-side policies that attempt to stop or reverse the wage increases. These could involve labour market measures such as lowering the minimum wage, or reducing the power of labour unions so that these are unable to negotiate high wage increases with employers.

If the increase in costs is due to an increase in the price of an imported input, then the solution is less obvious. An imported cause of cost-push inflation around the world over the past 40 years has involved increases in the price of oil, an input that is heavily used as energy in most lines of production in both industry and agriculture. There are no easy solutions to this type of cost-push inflation. Since the early 1970s, when the price of oil began to increase, many countries have attempted to address the problem through efforts to develop alternative forms of energy, as well as by encouraging users to economise on the use of products that depend on oil as an input. Such policies focus on reducing the *demand for oil*, so as to lower its price. If the price of oil falls, there results a rightward shift of the *SRAS* curve due to lower costs of production. However, this is a policy that takes a long time to take effect.

Another type of cost-push inflation may arise if firms with substantial market power (such as oligopolies) increase their profits by increasing the prices they charge to consumers. In this case, policies pursued may be to break up the market power of firms, and encourage competition (market-based supply-side policies).

Another type of cost-push inflation may occur if a country's currency falls in value, resulting in an increase in the prices it has to pay for imported goods (this will be explained in Chapter 16). Firms that are heavy users of imported inputs and raw materials experience an increase in their costs of production and cost-push inflation will result. One possible solution is to implement policies that aim to reduce dependence on imports ('expenditure switching' policies). However these policies come with their own problems, which we will discover in Chapter 17 (at HL).

Policies to promote economic growth

We must here make a distinction between short-term growth and long-term growth (see Chapter 11). Short-term growth is caused mainly by increases in aggregate demand. Long-term growth is caused by increases in long run aggregate supply (*LRAS*) or Keynesian *AS*, hence resulting in increases in potential output. Therefore the policies leading to short-term growth include expansionary fiscal and monetary policies, while the policies for long-term growth include supply-side policies, both interventionist and market-based.

However, it is important to consider the overlaps between demand-side and supply-side policies because in practice it is very difficult to distinguish what policy has what effect. Our theoretical distinctions between short-term and long-term growth become blurred in the real world. In fact, policy-makers do not usually make a distinction between the short term and long term or between expected effects on the demand side (*AD*) or the supply side (*LRAS*) of the economy, as they often focus just on the idea of growth in real GDP.

Therefore, when evaluating policies to achieve growth, we must remember that there is a wide variety of policies that include both expansionary demand-side policies as well as all supply-side policies, all of which can be expected to contribute to growth. All these policies have their advantages and disadvantages, and should be carefully selected by policy-makers in accordance with the particular circumstances of the country in question.

For example, if a country is in a deflationary gap and facing a recession with cyclical unemployment, appropriate policies are likely to be those discussed earlier in connection with cyclical unemployment. They include demand-side policies possibly along with supply-side policies, particularly of the interventionist type which actually overlap with expansionary fiscal policies. As the economy approaches potential output, or comes close to facing an inflationary gap, it becomes all the more important that supply-side policies are pursued, because demand-side policies on their own run the risk of creating inflationary pressures that can only be reduced by increases in the productive capacity of the economy, hence increases in *LRAS* or Keynesian *AS* (as Figure 11.3 in Chapter 11 shows).

Beyond the above, each type of policy must be evaluated on the basis of its own strengths and weaknesses.

TEST YOUR UNDERSTANDING 13.11

1. Answer the questions below for
   - a cyclical unemployment,
   - b structural unemployment,
   - c frictional and seasonal unemployment,
   - d demand-pull inflation,
   - e cost-push inflation.
     - i Identify policies you would recommend.
     - ii Identify what policies, if any, you would not recommend.
     - iii Explain the advantages and disadvantages of your policy recommendations.
2. Discuss why it is important to examine where in the business cycle an economy finds itself before recommending particular policies to achieve growth.

THEORY OF KNOWLEDGE 13.1

Paradigm shifts in macroeconomics

Thomas Kuhn was a physicist who became very well known for his work in the philosophy of science through his famous book, *The Structure of Scientific Revolutions*. Kuhn argued that science (and by extension, social science), does not grow and progress in a continuous way through a gradual build-up of knowledge, but rather progresses through abrupt 'scientific revolutions', known as paradigm shifts. The word *paradigm* comes from the Greek word παράδειγμα (*paradeigma*), which means pattern, or example, or representation. Kuhn used it to refer to a thought pattern that defines a scientific discipline at a particular time. According to Kuhn, a paradigm is not just a theory, but a whole world view that goes along with a theory or set of theories, which is shared by the members of the scientific community. A paradigm shift occurs when there is change in the paradigm of a discipline. A paradigm shift does not occur easily, as there is resistance to the shift by adherents to the paradigm being challenged.

Some economists argue that there occurred two major paradigm shifts in macroeconomics in the 20th century, and that there may be a third one occurring as a result of the global financial crisis that began in 2008.

CONTINUED

From classical economics to Keynesian economics

In the early 20th century, economists were guided in their macroeconomic policies by the principles of 'classical economics', which were based strongly on the microeconomic theory of supply and demand. Classical economists believed in the ability of the market and the price mechanism to solve all the major economic problems and allocate resources in the most efficient way. These principles, which are still accepted for the microeconomy, were then believed to apply to the macroeconomy as well. Major disruptions to the macroeconomy, affecting output and employment, were thought to be caused by factors external to the market system (such as wars, droughts or taxes), and were believed to be short-run phenomena that would be solved by the market, without interference by the government.

However, the Great Depression of the 1930s, which caused very large declines in output and large increases in unemployment, and which persisted for years, forced economists to question the ability of the market system to automatically generate the aggregate demand that was needed to get the economy out of the depression.

In 1936, John Maynard Keynes published his famous book, *The General Theory of Employment, Interest and Money*, in which he explained that the rigidity of wages and prices would not allow the market system to correct a recession (or depression) and bring the economy back to full employment. Soon after, with the outbreak of the Second World War in 1939, massive increases in government military expenditures showed economists the powerful effects of aggregate demand increases on employment and output.

The lessons of this experience brought forth a 'revolution' in economic thinking, or a paradigm shift, involving abandonment of the classical paradigm and a shift to the Keynesian one. In contrast to the classical world view, in which well-functioning markets meant there was not much for governments to do, in the Keynesian world view, markets were not self-correcting, and required active demand management (fiscal and monetary policies) to deal with economic fluctuations; government intervention in markets was indispensable for the proper functioning of the macroeconomy. By the end of the 1960s, many economists believed they had discovered in government intervention and demand management the secret to sustained economic growth, with low inflation and low unemployment.

From Keynesian economics to monetarist/new classical economics

In the early 1970s, this state of affairs was disrupted by the appearance of stagflation, or the simultaneous increase in inflation and unemployment, caused by falling aggregate supply (due to oil price and food price shocks). Economists realised that demand-side policies were not as effective as they were previously thought to be, since what was needed was expansionary policy for high unemployment and contractionary policy for inflation.

The 1970s saw high inflation and low output growth. This brought forth a new 'revolution' in economic thinking, or paradigm shift, involving the ideas of Milton Friedman, the founder of monetarism, which appeared to address the problem of stagflation, and also fitted in well with a political and ideological turn toward a market orientation that was occurring at the time as a reaction to the idea of big government (see Theory of knowledge 10.2 in Chapter 10). The world view of monetarist economists was similar to that of the classical economists, based on a belief in the ability of markets to lead to efficient outcomes and address the problems of the macroeconomy, again implying that the role of government should be small. Though demand management continued to be used, the new world view paved the way for the supply-side policies that dominated the 1980s and 1990s, which focused on trying to make the market system work more effectively.

The global financial crisis and a new paradigm shift?

The global financial crisis that began in 2008 revealed the weaknesses of excessive deregulation of the financial sector, and the failure of markets to work as well as had been supposed. In addition to the enormous sums of money spent by governments to save banks and other financial institutions from failing, there have also been numerous calls for increased regulation and oversight of financial activities on both national and global fronts. This raises the question whether another paradigm shift, involving

CONTINUED

greater government intervention and less reliance on the market may be imminent. Nobel Prize-winning economist, Joseph Stiglitz, writes:

'The blame game continues over who is responsible for the worst recession since the Great Depression . . . But the economics profession bears more than a little [responsibility]. It provided the models that gave comfort to regulators that markets could be self-regulated, that they were efficient and self-correcting . . . Today, not only is our economy in a shambles but so too is the economic paradigm that predominated in the years before the crisis . . .

Fortunately, while much of the mainstream focused on these flawed models, numerous researchers were engaged in developing alternative approaches. Economic theory had already shown that many of the central conclusions of the standard model were not [reliable] . . .

Changing paradigms is not easy. Too many have invested too much in the wrong models. Like the Ptolemaic attempts to preserve earth-centric views of the universe, there will be heroic efforts to add complexities and refinements to the standard paradigm. The resulting models will be an improvement and policies based on them may do better, but they too are likely to fail. Nothing less than a paradigm shift will do.

But a new paradigm, I believe, is within our grasp: the intellectual building blocks are there . . .' [8]

Joseph Stiglitz and another Nobel Prize-winning economist, George Akerlof, note:

'The economic and financial crisis has been a telling moment for the economics profession, for it has put many long-standing ideas to the test. If science is defined by its ability to forecast the future, the failure of much of the economics profession to see the crisis coming should be a cause of great concern . . .

Just as the crisis has reinvigorated thinking about the need for regulation, so it has given new impetus to the exploration of alternative strands of thought that would provide better insights into how our complex economic system functions . . .

Fortunately, while some economists were pushing the idea of self-regulating, fully efficient markets that always remain at full employment, other economists and social scientists have been exploring a variety of different approaches . . .

Much of the most exciting work in economics now under way extends the boundary of economics to include work by psychologists, political scientists, and sociologists. We have much to learn, too, from economic history.'[9]

Thinking points

- Can you detect a pattern in the balance between markets and government intervention that has been occurring in the shift from one paradigm to another?
- Can you think of any paradigm shifts that may have occurred in another social science or science you are studying?
- Based on the paradigm shifts described above, what has happened in the economy in each case to bring forth a paradigm shift?
- What kind of events do you think are likely to lead to paradigm shifts in other social sciences and in the natural sciences? Do they differ from those in economics?
- Why do you think paradigm shifts do not happen easily? Why do they occur infrequently?

[8] Joseph Stiglitz, 'Needed: a new economic paradigm' in the *Financial Times*, 19 August 2010.

[9] George Akerlof and Joseph E. Stiglitz, 'Let a hundred theories bloom' in *Project Syndicate*, 26 October 2009.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Select a country of your choice that is or has recently been in recession and examine what policies the government or central bank have pursued in order to deal with this. Consider the extent to which the policies have been successful as well as the difficulties that the policies may have encountered in attempting to achieve their goals.

2 Find four recent news sources with examples of the following four kinds of macroeconomic policies:

   a fiscal policy,

   b monetary policy,

   c interventionist supply-side policy, and

   d market-based supply-side policy.

For each one of these, describe the policy. For example, in the case of fiscal policy is it contractionary or expansionary; does it involve changes in taxes or government spending; what is the objective of the policy. In addition, you should note if any of the policies fall into more than one category. For example, some fiscal policies overlap with interventionist or market-based supply-side policies and vice versa.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Unit 4

The Global Economy

In our study of economics so far, we have been examining 'closed' economies, or economies that are closed to economic relations with other countries. In the real world, countries have many economic links with other countries, involving flows of goods, services and resources, as well as payments and money across countries for many purposes. These international economic links, which make countries more and more interdependent, are known as **economic integration**. Chapters 14–17 are concerned with these international economic links.

Later in this unit we will examine the special problems of developing countries. In Chapters 18–20 we will use the principles developed up to that point to understand the circumstances of developing countries in the global economy and what can be done to hasten their development and improve standards of living and well-being everywhere.

Real world issue 1: Who wins and who loses from the integration of the economies of our world?

CONCEPTUAL UNDERSTANDINGS

1. Growing economic ***interdependence*** between countries brings with it benefits as well as costs.
2. Growing economic integration may lead to greater ***efficiency***, welfare gains and improvements in ***economic well-being***, but the benefits may be ***inequitably*** distributed.

In Chapters 14 and 15 we will study international trade *theory*, explaining why countries trade with each other, and what are the benefits of trade. We will also examine international trade *policy*, where we will discover the reasons why countries use policies or measures that affect international trade, sometimes restricting it, as well as the consequences of these policies.

Chapters 16–17 will be concerned with flows of payments from one country to another. We will see how countries measure flows of money they receive from other countries and that they send out to other countries. We will also study the problems that can arise in the event of imbalances between these flows, as well as policies to correct such imbalances. We will learn that there are strong interconnections between events in the domestic economy and the international economy, which mean that often economic policies to correct a domestic economic problem have impacts on economic relations with other countries.

In our study of international economics, we will discover that the benefits of economic integration are unevenly distributed: some countries or groups within countries gain while others may be worse off. We will examine who gains, who loses and why.

Reconstruction of a classical era Greek ship, watercolour, France, 19th century

Chapter 14

International trade: Part I

BEFORE YOU START

- Can you think of some reasons why countries trade with each other?
- You have probably heard news reports of trade wars. Can you think of some reasons why countries might want to restrict trade?

International trade, the buying and selling of goods and services across international boundaries, has taken place since ancient times, by Egyptians, Greeks, Romans and Phoenicians, and later by all major powers throughout history up to the present. But it has never been so important to the economies of virtually all countries in the world as it is today. Goods and services produced domestically sold to buyers abroad are *exports*; goods and services bought from other countries for domestic use are *imports*. In recent decades, the value of exports and imports as a share of GDP has been increasing in most countries around the world.

You may recall from your study of macroeconomics that net exports (= exports minus imports, or $X - M$), is one of the components of aggregate demand. It is a component with major consequences for domestic economies as well as the global economy, some positive and some negative. On the whole the positive consequences are often greater than the negative ones, and yet there are many reasons why countries sometimes try to restrict the flow of trade. These restrictions often give rise to much disagreement and even conflict within and between countries.

In this chapter we will study the benefits of international trade, as well as different kinds of measures used by countries to restrict the amount of trade, known as *trade protection*.

Photo: Hong Kong. Container freight ship terminal

14.1 The benefits of international trade

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the various benefits that countries can expect from trade (AO2)
- draw diagrams illustrating free trade resulting in exports or in imports (AO4)
- calculate, from free trade diagrams, quantity of exports, export revenue; quantity of imports, import expenditure (HL only) (AO4)

Why do countries benefit from international trade?

International trade results in a number of important benefits or **gains from trade**. The benefits are far greater if trade is free. **Free trade** refers to the absence of government intervention of any kind in international trade, so that trade takes place without any restrictions (barriers) between individuals, firms or governments of different countries.

Increased competition

When countries trade with each other, they import products from other countries so that domestic firms become exposed to competition from products produced by firms in foreign countries. In addition, they export products to other countries so their own products are forced to compete with other products produced in those countries. Greater competition results in several benefits discussed below.

Greater efficiency in production

As a result of greater competition firms are forced to become more efficient, in other words they must try to produce at the lowest possible cost. If they do not become more efficient, they will have to sell their output at higher prices to cover their higher costs; domestic consumers will prefer the lower-priced imported products, while foreign consumers will prefer lower-priced goods produced in other countries. As a result, higher-cost firms will have lower sales and may go out of business. Therefore, increased competition leads to greater efficiency.

Lower prices for consumers

Increased competition among firms and greater efficiency lead to lower prices for consumers. In addition, as imports consist of goods that are produced more efficiently in other countries, this also leads to lower prices for consumers.

Greater choice for consumers

By trading with each other, countries can import a larger variety of goods and services, possibly of higher quality, than the ones they can produce themselves. This increases choice for consumers.

Acquiring needed resources

Countries are likely to need for their domestic production a variety of natural resources or capital goods that are not available domestically. For example, oil is a resource that virtually all countries depend on, yet most are forced to rely on imported oil because they do not produce it themselves. The same may apply to a variety of other resources such as timber, minerals and semi-finished products used as inputs, as well as capital goods (machinery and equipment) used in production. Trade allows countries to import inputs they need for domestic production.

Source of foreign exchange

When countries export, they acquire foreign exchange (or foreign currencies), which allows them to make payments to other countries for the goods and services they import, or make other payments abroad (see Chapter 16). Acquiring foreign exchange from exports increases their ability to import.

Access to larger markets

In the absence of trade, the amount of output any firm can produce is limited by the size of the domestic market. If the domestic market is small (if the country is small), the firm is unable to grow as it does not have a market for increasing its sales. With trade and exports to other countries there may be an expansion in the size of the market and therefore the possibility of expanding sales of the firm.

Economies of scale in production

Economies of scale involve the ability of firms to decrease average costs of production (cost per unit of output) by becoming larger and increasing the quantity of output produced. (This topic was studied in Chapter 7 at HL.) When a firm lowers its average costs, it becomes more efficient, and can sell its output at a lower price. Access to larger markets allows firms to grow beyond the limits of national boundaries, produce more output and take advantage of economies of scale. With lower average costs, they can lower their prices and enjoy greater export competitiveness, or the ability to compete better and sell more in foreign markets.

Increases in domestic production and consumption as a result of specialisation

Many of the benefits of trade arise from **specialisation**. Specialisation occurs when an individual, firm or country concentrates production on one or a few goods and services. Here, we are referring to specialisation by a country in the production of goods or services it can produce efficiently (at a low cost). A country that does not trade must itself produce all the goods and services consumed, and therefore cannot specialise. But if it uses its resources to specialise in the production of those goods and services it can produce more efficiently (with lower costs of production, or with fewer resources), it can produce more of these, and trade some of them for other goods produced more efficiently in other countries. This way it is able to produce a greater quantity of output because it does not 'waste' its scarce resources on producing goods and services at a relatively high cost. It can also increase its consumption of goods and services, because by exporting part of its larger domestic output in exchange for other output produced more cheaply elsewhere, it can acquire a larger overall quantity of goods and services. This, in a simple form, is the theory of comparative advantage that we will study below.

More efficient allocation of resources

If trade is free, meaning there are no restrictions on trade, it can lead to a more efficient allocation of resources both within countries as well as globally. This follows from specialisation discussed above. If each country specialises in producing the goods it can produce relatively more efficiently with lower costs of production or fewer resources compared with its trading partners, there will be less waste of scarce resources and therefore more efficient resource allocation.

Trade makes possible the flow of new ideas and technology

As goods and services flow from one country to another, they enable new ideas and new technologies and skills to be transferred from one country to another.

Trade makes countries interdependent, reducing the possibility of hostilities and violence

Strong international trade links between countries can form the basis for economic relationships that reduce the possibility of war or other hostilities. One of the reasons behind the establishment of the European Economic Community in 1957 (the EEC, the precursor of the European Union) was to eliminate the possibility of future wars between France and Germany. The strong economic interdependence created by trade (and other) links between these countries makes the possibility of war between them inconceivable today.

Trade as an 'engine for growth'

Increased competition, greater efficiencies in production, expanding markets, acquisition of needed resources, economies of scale, greater specialisation, improved resource allocation and sharing of technological advances, all made possible by international trade, contribute to increases in domestic output, and therefore to greater economic growth. For these reasons international trade has been termed an 'engine for growth'.

TEST YOUR UNDERSTANDING 14.1

1. Identify what products or groups of products your country specialises in.
2. Discuss how each of the following can benefit from trade:
   - a consumers
   - b producers
   - c the domestic economy and society
   - d the global economy and society
3. Outline why international trade been termed an 'engine for growth'.

Using diagrams to illustrate free trade

We have seen that *free trade* refers to the absence of government intervention of any kind in international trade, resulting in trade without any restrictions. Under free trade the prices of goods that are traded internationally (imported and exported) are determined entirely by the forces of demand and supply.

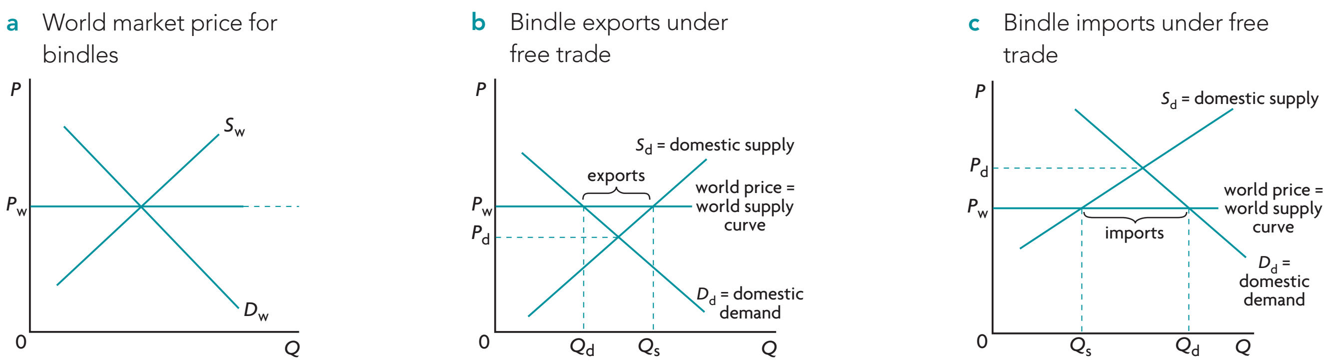


Figure 14.1: Using diagrams to illustrate free trade

Should a country export or import a good?

Suppose that bindles are a good produced by many countries around the world, which have free trade for bindles. This means that there is a free world market for bindles, involving many individuals or firms in countries around the world who buy and sell bindles. The world bindle price is determined by world demand and world supply, where world demand is the sum of all country demands and world supply is the sum of all country supplies. All the countries that are part of this world market buy and sell bindles at the world price. The world price, P_w, is shown in Figure 14.1(a).

Imagine now a country called Tradenia that also produces bindles but is closed to international trade; this is called *autarky*, the Greek word for *self-sufficiency*, since the country is self-sufficient in all the goods it produces and consumes. Tradenia is not part of the world market so its domestic price of bindles is determined entirely by the familiar domestic demand and domestic supply studied in Chapter 2.

Tradenia then decides to open its economy to international trade, so that it will now begin to export and import various goods. In the case of bindles, the question arises, should these be exported or imported? *The answer depends on the domestic price of bindles before trade, compared with the price of bindles in the world bindle market*.

It is assumed (for simplicity) that the Tradenian bindle market is relatively small compared to the overall size of the world bindle market, so that if Tradenia buys (imports) or sells (exports) bindles in the world market it will not influence the world bindle price. This means that once Tradenia enters the world bindle market it will accept the world price of bindles so that its previous domestic price before trade no longer exists. In other words, Tradenia now faces a perfectly elastic bindle supply curve, appearing as a horizontal line at the level of the world price (see Chapter 3 on elasticities). *A perfectly elastic supply curve at the world price simply means that all bindles in Tradenia are bought and sold at the world price and no other price*. [1]

When a country should export a good

Whether Tradenia should export or import bindles is explained in Figure 14.1. Figure 14.1(b) shows the case where Tradenia becomes a bindle exporter. The world price, P_w, determined in the world bindle market, is higher than Tradenia's domestic price, P_d. Once Tradenia opens its economy to international trade and joins the world market, it accepts the world price P_w, and the domestic price, P_d is no longer relevant. At the higher price P_w, the quantity of bindles supplied, Q_s, is larger than the quantity of bindles demanded, Q_d. This excess quantity supplied, which is $Q_s - Q_d$, is available to be sold to buyers abroad, or exported. It follows then that *under free trade, when the world price is higher than the domestic price, the good in question is exported.*

When a country should import a good

Figure 14.1(c) shows that the world price of bindles determined in the world market, P_w, is lower than Tradenia's domestic price, P_d. At the world price P_w, the quantity of bindles demanded, Q_d, is larger than quantity of bindles supplied, Q_s. Tradenia now has an excess quantity demanded, $Q_d - Q_s$, which is the quantity of bindles to be purchased from abroad, or imported. *Under free trade, when the world price is lower than the domestic price, the good in question is imported.*

[1] Students taking this course at HL may note that this is similar to firms in perfect competition that are price-takers, accepting the price determined in the market and selling all they can at that price. See Chapter 7.

A country will export a good if its domestic price without trade is lower than the world price and it will import a good if its domestic price without trade is higher than the world price.

TEST YOUR UNDERSTANDING 14.2

Using international trade diagrams, explain when a country should export a good and when it should import a good.

Calculating quantities of exports or imports, and export revenues or import expenditures under free trade (HL only)

We will use the principles explained above to calculate the effects of international trade on several variables, based on the information presented in Figure 14.2.

The exporting country

Figure 14.2(a) shows that when Tradenia was a closed economy, it was producing 60 million bindles and selling them at a price of $5 per bindle. When it opened up its economy to international trade, it began accepting the global price of $7 per bindle.

Greater quantity produced, smaller quantity consumed and quantity of exports

At this higher price, domestic quantity produced increased from 60 million to 100 million bindles, while domestic quantity demanded fell from 60 million to 20 million bindles. This means that there is now excess quantity supplied of 100 million – 20 million = 80 million bindles. *This excess quantity supplied will be exported therefore exports = 80 million bindles*.

Export revenues

As a result, domestic producers will now be earning export revenues which are equal to the quantity of exports times the world price they receive per unit. *Export revenues are therefore 80 million x $7 = $560 million*.

Who wins and who loses

Producers benefit on two counts: the higher price as well as the greater quantity they produce which means that they have higher revenues. Consumers on the other hand are worse off on two counts: they can only buy a smaller quantity and must pay a higher price.

The importing country

Figure 14.2(b) shows that the world price that Tradenia is now facing is $3 per bindle, which is lower than the domestic price of $5 without trade.

Smaller quantity produced, greater quantity consumed and quantity of imports

At this lower world price, the domestic quantity produced fell from 60 million to 20 million bindles. At the same time, domestic quantity demanded increased from

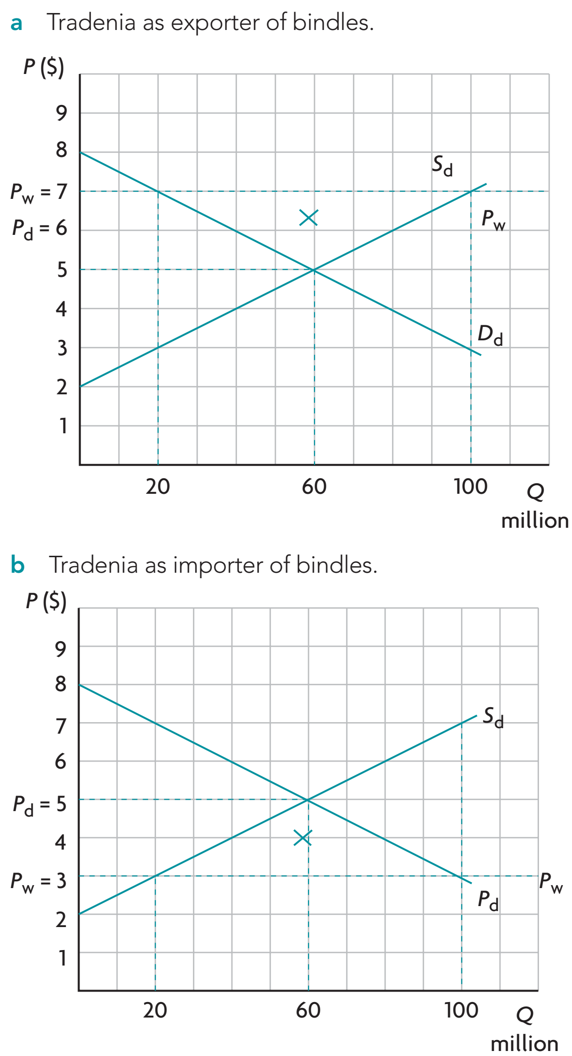


Figure 14.2: Exporting or importing under free trade

60 million to 100 million bindles. There is therefore excess quantity demanded of 100 million – 20 million = 80 million bindles. *This excess quantity demanded will be satisfied by imports therefore imports = 80 million bindles.*

Import expenditure

These imports will give rise to an amount of import expenditure (the amount of spending to buy imports) that is equal to the quantity of imports times the world price per unit. *Therefore import expenditure is 80 million x $3 = $240 million.*

Who wins and who loses

Producers are now worse off because they produce a smaller quantity which they sell at a lower price. Consumers on the other hand are better off since they can buy a larger quantity for a lower price.

TEST YOUR UNDERSTANDING 14.3

1 Tradenia's domestic prices before trade are $50 per unit of good X and $200 per unit of good Y. When it begins to trade it faces a world price of $35 for good X and $225 for good Y.

- a Identify which good Tradenia will export and which it will import. Explain why.
- b Identify who will gain or lose from importing a good and who will gain or lose from exporting a good.

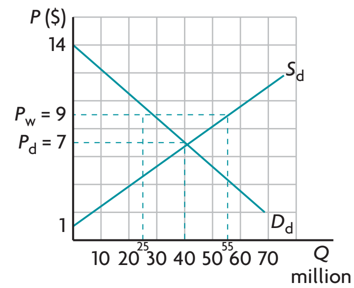
Calculating effects on producers, consumers and social surplus (Supplementary material)

If you are interested in pursuing this topic further you may do so by referring to the digital coursebook. You will discover that welfare analysis reveals the interesting result that free trade always increases social surplus regardless whether the country is an importer or exporter, though the effects are unevenly shared between consumers and producers.

TEST YOUR UNDERSTANDING 14.4

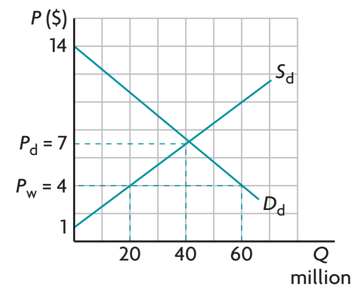
1 Using the diagram below, showing Tradenia before trade and after trade, calculate

- a the quantity of exports,
- b export revenues,
- c (optional*) the change in producer revenue,
- d (optional*) change in consumer expenditure.



2 Using the diagram below, showing Tradenia before and after trade, calculate

- a the quantity of imports,
- b import expenditures,
- c (optional*) the change in producer revenue, and
- d (optional*) the change in consumer expenditure.



*This is not required by the syllabus but it is good practice for later problems

14.2 Free trade: absolute and comparative advantage (HL only)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the theory of absolute advantage using the concept of gains from trade (AO2)
- explain the theory of comparative advantage using the concepts of gains from trade and opportunity cost (AO2)
- explain the sources of comparative advantage (AO2)
- evaluate the theory of comparative advantage (AO3)
- draw a linear *PPC* diagram illustrating different opportunity costs, and gains from specialisation and trade arising from comparative advantage (AO4)
- calculate opportunity costs from data and identify comparative advantage. (AO4)

The theories of absolute and comparative advantage provide powerful explanations for the principle that countries can gain significant benefits through specialisation and trade.

The theory of absolute advantage

The theory of absolute advantage dates back to the work of the famous 18th century economist Adam Smith, considered to be the 'Father of Economics' (see Chapter 1). Adam Smith was a strong believer in the principle that specialisation and free trade can make all countries better off. To explain this, he formulated the theory of absolute advantage. **Absolute advantage** refers to the ability of one country to produce a good using fewer resources than another country. Putting it differently, *a country has an absolute advantage in a good if with the same quantity of resources it can produce more of the good than another country*.

Consider a simple world economy of two countries, Coffenia and Robotia, that produce coffee and robots. In Table 14.1, columns 1 and 2 show the quantities of coffee and robots that one worker in one day can produce in Coffenia and in Robotia, if they produce only coffee or only robots. Coffenia can produce either 8 units of coffee (and 0 robots), or it can produce 4 robots (and 0 units of coffee). Similarly Robotia can produce either 3 units of coffee (and 0 robots) or 6 robots (and 0 units of coffee).

Coffenia therefore has an absolute advantage in coffee, because with one worker it can produce 8 units of coffee compared to only 3 in Robotia. Robotia has an absolute advantage in robots, since it can produce 6 robots, compared to only 4 in Coffenia.

	Production possibilities when each country produces only coffee or only robots		
	1 Coffee		2 Robots
Coffenia	8	or	4
Robotia	3	or	6

Table 14.1: Absolute advantage

Using this information, we can construct production possibilities curves (*PPC*s) for Coffenia and Robotia, shown in Figure 14.3. For simplicity, we use straight-line *PPC*s (see Chapter 1). When Coffenia produces 8 units of coffee, it is at point A, producing 0 robots; and when it produces 4 robots, it is at point B producing 0 units of coffee. In the same way we plot points C and D for Robotia. Joining points A and B, we get Coffenia's *PPC*; joining points C and D gives Robotia's *PPC*.

Comparing the two *PPC*'s, we immediately see Robotia's absolute advantage is in robots, because its *PPC* extends further to the right on the robot axis; and Coffenia's absolute advantage is in coffee, since its *PPC* extends further up on the coffee axis.

Suppose that both countries agree to specialise in and export the good in which they have absolute advantage. Coffenia specialises entirely in coffee production, and exports part of its coffee in exchange for robot imports. Robotia specialises entirely in robot production, and it exports a portion of its robots in exchange for coffee

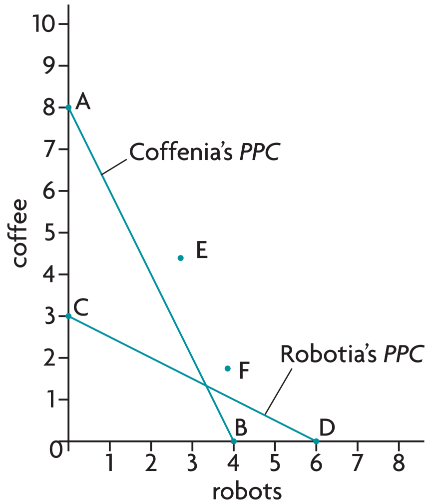


Figure 14.3: Coffenia has absolute advantage in coffee; Robotia has absolute advantage in robots

imports. The result will be that whereas both countries will be producing somewhere on their *PPC*, *due to specialisation and trade they can consume at a point outside their PPC!* This is shown in Figure 14.3. Both countries become better off because specialisation according to absolute advantage leads to a 'global' reallocation of resources where production takes place by the most efficient (low-cost) producers.

According to the **theory of absolute advantage**, if countries specialise in and export the good in which they have an absolute advantage (can produce with fewer resources) the result is increased production and consumption in each country.

The theory of comparative advantage

The theory of absolute advantage explains only a small part of gains from specialisation and trade. A much more powerful argument was provided by a well-known economist of the 19th century, David Ricardo, in his famous theory of comparative advantage. Ricardo showed that countries can gain from specialisation and trade even if one country has an absolute advantage in *both goods*. In order for this surprising result to hold, it is only necessary that countries have *different opportunity costs* for their goods, so that the production of one good is relatively cheaper to produce in one country than in another, even if it is not absolutely cheaper. **Comparative advantage** refers to the situation where one country has a lower opportunity cost (relative cost) in the production of a good than another country.

Using diagrams to show comparative advantage

Consider a simple world economy of two countries, Cottonia and Microchippia, producing cotton and microchips. Table 14.2 shows the quantities of each good that one worker can produce in one day if only one or the other good is produced. Cottonia can produce either 20 units of cotton (and zero units of microchips) or it can produce 10 units of microchips (and zero cotton). Microchippia can produce either 25 units of cotton (and zero units of microchips) or 50 units of microchips (and zero units of cotton). We can see that Microchippia has an absolute advantage in the production of *both cotton and microchips*, because with the same resources (one worker in one day) it can produce more of both goods than Cottonia.

Figure 14.4 plots the *PPC*s of each of the two countries based on the data of Table 14.2 (assuming straight-line *PPC*s). Microchippia's absolute advantage in the production of both goods is apparent from the fact that *its PPC lies entirely above the PPC of Cottonia*. When comparing the PPCs of two countries, we can see immediately whether one country has the absolute advantage in one or both of the goods. If the *PPC*s intersect, as in Figure 14.3, this means that each country has an absolute advantage in one of the two goods. If the *PPC*s do not intersect, as in Figure 14.4 this means that the country with the *PPC* lying fully above the second *PPC* has an absolute advantage in the production of both goods (in our example, Microchippia).

If the two *PPC*s do not intersect, as in Figure 14.4, how can we determine comparative advantages? Very simply, *the country that has the flatter PPC has a comparative advantage in the good measured on the horizontal axis*. It follows that the country with the steeper *PPC* has a comparative advantage in the good measured along the vertical axis. In Figure 14.4, Microchippia's *PPC* is flatter than Cottonia's *PPC*, therefore Microchippia has a comparative advantage in microchips, which are measured along the horizontal axis. Cottonia, with the steeper *PPC*, has the comparative advantage in cotton.

Note that when a country does not have a comparative advantage in the production of a good, we can say it has a *comparative disadvantage*. Therefore in our example above, Cottonia has a comparative disadvantage in the production of microchips..

	Production possibilities when each country produces only cotton or only microchips			Opportunity cost of cotton	Opportunity cost of microchips
	1 Cotton		2 Microchips	3	4
Cottonia	20	or	10	$\frac{\text{10 units of microchips}}{\text{20 units of cotton}} = \frac{1}{2}$	$\frac{\text{20 units of cotton}}{\text{10 units of microchips}} = 2$
Microchippia	25	or	50	$\frac{\text{50 units of microchips}}{\text{25 units of cotton}} = 2$	$\frac{\text{25 units of cotton}}{\text{50 units of microchips}} = \frac{1}{2}$

Table 14.2: Comparative advantage

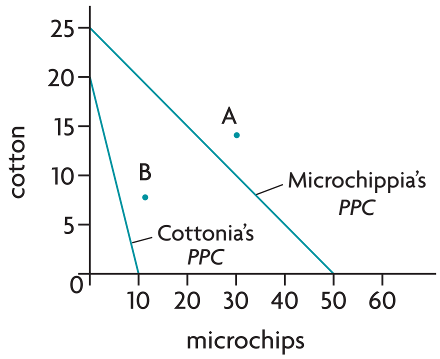


Figure 14.4: Comparative advantage

Calculating opportunity costs from a set of data

We will now calculate opportunity costs to identify comparative advantage. Opportunity cost is the next best alternative that must be sacrificed in order to obtain something (see Chapter 1).

Columns 3 and 4 in Table 14.2 calculate the opportunity costs of the two goods in each country. (Since we are using straight-line *PPC*s, this means that opportunity costs are constant throughout the *PPC*). In general you can use the following rule to calculate opportunity cost:

$$\text{Opportunity cost} = \frac{\text{sacrifice of one good}}{\text{gain of the other good}}$$

Therefore for Cottonia, the opportunity cost of cotton is the quantity of microchips that must be sacrificed (10 units) divided by the gain in cotton (20 units). The opportunity cost of microchips is the amount of cotton that must be sacrificed (20 units) divided by the gain in microchips (10 units). We perform similar calculations to find the opportunity costs for Microchippia.

The results show that Microchippia has a lower opportunity cost in producing microchips, while Cottonia has a lower opportunity cost in producing cotton. Though Cottonia has a *higher absolute cost* in producing cotton, it has a *lower relative cost*, meaning that if Cottonia wants to produce more cotton, it needs to sacrifice a smaller quantity of microchips than does Microchippia. *Therefore Cottonia has a comparative advantage in cotton production, while Microchippia has comparative advantage in microchips.*

This ties in with our conclusions above based on the *PPC*s of Cottonia and Microchippia. In fact, it is possible to calculate opportunity costs directly from Figure 14.2.

A country has a comparative advantage in the production of the good that has a lower opportunity cost (lower relative cost).

The theory of comparative advantage

The theory of comparative advantage is so important that it is often referred to as the *law of comparative advantage*. This states that if countries specialise and trade according to their comparative advantage, global production and consumption will increase because of an improvement in the global allocation of resources, making all countries involved better off.

Because of this improvement in resource allocation, even though both countries produce at a point on their *PPC*s, *through specialisation and trade they consume at a point outside their PPC!* Cottonia can consume at a point like B while Microchippia can consume at a point like A in Figure 14.4.

A country has a comparative advantage in the production of a good when this can be produced at a lower opportunity cost. According to the **theory (or law) of comparative advantage**, as long as opportunity costs in two (or more) countries differ, it is possible for all countries to gain from specialisation and trade according to their comparative advantage. The global allocation of resources improves, resulting in greater global output and greater global consumption, allowing countries to consume outside their *PPC*.

The case of parallel PPCs

What if the *PPC*s of two countries are parallel to each other, as in Figure 14.5? Here, country A has an absolute advantage in the production of both good Y and good X. The fact that the two *PPC*s are parallel means that the two countries face identical opportunity costs for the two goods.[3] If opportunity costs are identical, there is no country in which one good is relatively cheaper; therefore there is no country that has a comparative advantage in the production of one or the other good. Under these circumstances (which do not occur very often in the real world), there are no possibilities for countries to gain from specialisation and trade, and there would be no point in trading.

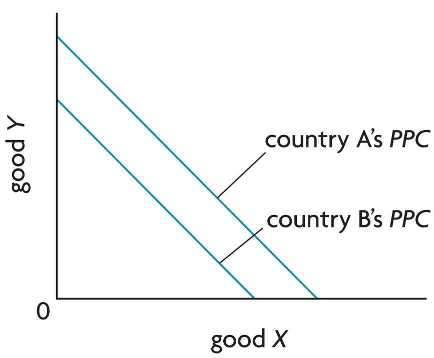


Figure 14.5: Identical opportunity costs: no gains from trade

Absolute advantage as a special case of comparative advantage

As we have seen, according to the theory of comparative advantage, countries gain from specialisation and trade as long as they have different relative costs of production. This means that the theory of absolute advantage is simply a special case of the theory of comparative advantage. Considering the case of Coffenia and Robotia, the fact that Coffenia has an absolute advantage in coffee and Robotia in robots does not affect the conclusions of the theory of comparative advantage.

Revisiting the question 'Should a country export or import a good?'

Now that we have studied the theory of comparative advantage, we can better understand the diagrams presented in Figure 14.1, which in fact are illustrations of the principle of comparative advantage. In Figure 14.1(b), the lower domestic price compared to the world price indicates that Tradenia has a *comparative advantage* in the production of bindles, it can produce these more efficiently (at a lower cost) than other countries, and can therefore sell them at a lower price. When it opens the economy to international trade, it accepts the higher world price, increases its domestic production of bindles and exports some of these to other, less efficient countries. In Figure 14.1(c), Tradenia's higher domestic price indicates that it has *a comparative disadvantage*: it is less efficient than other countries. When it opens itself to trade, it accepts the lower world price, and the excess demand is satisfied by imports produced in other countries more efficiently.

The sources of comparative advantage

The possibility of increased production and consumption through specialisation is made possible because countries can make use of differences in quantities and quality of factors of production, as well as levels of technology, which altogether are called 'factor endowments'. For example, a country with a temperate climate will find it more costly to produce crops such as coffee or cocoa, which are better suited to tropical climates. Mountainous countries are less well suited to agriculture than countries with fertile plains.

Depending on their factor endowments, different countries are more efficient in the production of certain goods and services than others. Greece is a mountainous country

[3] This follows from the point that two parallel lines have identical slopes, and since the slope is the opportunity cost of the good measured on the horizontal axis, it follows that opportunity costs are identical.

with a large coastline, countless beaches and rich historical sites. It is therefore better suited to producing shipping services as well as tourism services. Switzerland, being a landlocked country, is not well suited to shipping, but has developed technologies that have made it well suited to the production of high-quality watches and clocks.

Strengths and limitations of the theory of comparative advantage

The theory of comparative advantage forms the basis of trade policies in many countries. Its key conclusion, that free trade increases global production and consumption, leading to an improved global allocation of resources, forms the justification of the major policy trend since the early 1990s around the world toward **trade liberalisation**, involving the freeing up of trade through the gradual removal of trade restrictions. However, in spite of its potentials, the theory of comparative advantage is subject to several limitations:

The theory of comparative advantage depends on many unrealistic assumptions:

- **Factors of production are assumed to be *fixed***, in other words they do not move from one country to another and do not change. Yet in the real world, factors of production, particularly labour and capital, can and often do move from country to country; moreover, there are likely to be changes in quality, such as when labour acquires more skills and education. This means that factor endowments change over time, and so comparative advantage also changes.
- **Technology is assumed to be *fixed***; this is highly unrealistic since new technologies are continuously being introduced. This too causes comparative advantage to change.
- **There is perfect mobility of factors of production within the country**, meaning that factors of production can be instantly and costlessly moved from one line of production to another. This does not occur in practice, as there are costs when production switches from one product to another that may be so high as to change comparative advantage.
- **There is full employment of all resources** (countries produce *on their PPCs*); this is hardly ever met, especially in developing countries, where there is often very high unemployment and underemployment of labour. When this occurs, comparative advantage is different from what it would have been with full employment.
- **There is free trade**, meaning that trade flows (imports and exports) are determined entirely by market forces; in reality, there is strong government intervention in markets that influences quantities of imports and exports.
- **There are homogeneous products**; the product is identical in every respect, for example all computers are exactly the same. This is rarely if ever the case in the real world.
- **Transportation costs are ignored**. In the real world, there are costs of transportation for imports and exports that change relative prices and may limit the benefits of specialisation.

Specialisation according to comparative advantage may not allow necessary structural changes to occur in an economy As an economy grows and develops, major changes in its structure usually occur, with the agricultural sector becoming less important, and manufacturing and services becoming more important (see Figure 3.12, Chapter 3). These changes are especially important for developing countries, indicating that comparative advantage changes over time. For example, an agricultural economy may have a comparative advantage in agricultural products, but as it becomes more industrialised, its comparative advantage may change in favour of manufactured products. If countries specialise according to their comparative advantage, they would have to go on producing and exporting according to that same advantage, and this would not permit the necessary structural changes to take place in the economy. This is an important issue for developing countries that we will come back to later in this chapter and in Chapter 19.

Trade on the basis of comparative advantage may lead to excessive specialisation If a country has a comparative advantage in only one or a few products, specialisation according to comparative advantage may lead to too much specialisation, which may make countries vulnerable if they become too dependent on exports of these products. For example, if there is a fall in exports due to a global recession, or a fall in export prices due to declining demand, there will result falling revenues from exports, falling incomes and economic decline. Further, primary products (including agricultural products) are subject to strong price fluctuations, which lead to unstable export revenues, also with negative impacts on the economy (see Chapter 19).

TEST YOUR UNDERSTANDING 14.5

1 a Explain the difference between 'absolute advantage' and 'comparative advantage'.

b Outline which of the two concepts is a more powerful explanation of the benefits from trade.

2 a Draw a diagram showing Oceanland's absolute advantage in shipping services and Flatland's absolute advantage in agricultural products.

b Using PPC diagrams, explain how Oceanland and Flatland can gain from specialisation and trade.

3 Lakeland has an absolute advantage over Mountainland in the production of both fish and computers, but Lakeland has a comparative advantage in fish production.

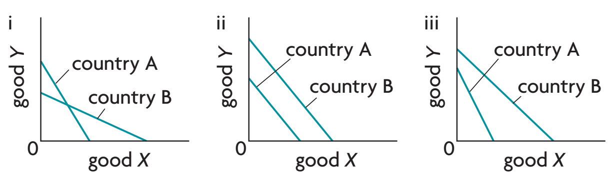
a Draw a diagram showing the absolute and comparative advantages of the two countries.

b Explain how Lakeland and Mountainland can gain from specialisation and trade.

4 a Using the data in Table 14.1, calculate the opportunity cost of coffee and robots in Coffenia and Robotia.

b Use your results of differing opportunity costs to explain why the theory of absolute advantage is a special case of the theory of comparative advantage.

5 Answer these questions based on the following diagrams:



a In diagram (i), identify which of the two countries has the absolute advantage in the production of which goods.

b In diagram (ii), can either country benefit from specialisation and trade? Why or why not?

c In diagram (iii), does any country have an absolute advantage in the production of either good?

d In diagram (iii), can either country benefit from specialisation and trade? Why or why not? If they can benefit, what good should each one specialise in?

6 According to the theory of comparative advantage, outline under what circumstances it is not worthwhile for countries to specialise and trade.

7 For each case below,

a calculate opportunity costs for Country A and Country B, and determine the good in which each country has a comparative advantage (if any),

b draw *PPC* diagrams indicating comparative advantage.

c indicate which good each country should specialise, export and import.

Production possibilities for Country A and Country B

	Country A	Country B
good X	8	2
good Y	2	4
good X	8	2
good Y	6	4
good X	1	4
good Y	2	2
good X	6	3
good Y	3	1
good X	1	2
good Y	2	4

8 a Outline some of the unrealistic assumptions on which the theory of comparative advantage rests.

b Explain some of the problems that countries may run into if they specialise and trade according to their comparative advantage.

14.3 Types of trade protection: restrictions on free trade

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- for each of the following types of trade protection (i) tariffs, (ii) quotas, (iii) subsidies, (iv) export subsidies:
  - explain the effects on markets and stakeholders (AO2)
  - draw diagrams illustrating effects on price, production, consumption, expenditures, revenues, welfare (AO4)
  - calculate, from diagrams, effects on stakeholders (HL only) (AO4)
- explain administrative barriers (AO2)
- discuss advantages and disadvantages of (AO3)
  - tariffs
  - quotas
  - subsidies
  - export subsidies
  - administrative barriers

Free trade versus trade protection

The benefits of trade and particularly free trade were discussed at the beginning of this chapter. In contrast to free trade, **trade protection** involves government intervention in international trade through the imposition of trade restrictions (barriers) to prevent the free entry of imports into a country. This is done to protect the domestic economy, particularly domestic firms and their workers, from foreign competition.

As Figure 14.1(c) shows, under free trade a good is imported when its domestic price is higher than the world price, indicating that a country has a comparative disadvantage (i.e. is relatively inefficient) in producing the good. We saw that when the country begins to trade, producers become worse off because they are forced to receive the lower world price and so produce a smaller quantity, resulting in reduced producer revenues. This is the reason why producers and their workers often favour trade protection measures that will limit the quantity of imports and make them better off.

The topic of free trade versus trade protection is highly controversial and has occupied economists for over 300 years. In recent decades it has become one of the most important international policy issues.

Tariffs

Tariffs, also known as 'customs duties', are taxes on imported goods, and are the most common form of trade restriction. Tariffs may serve two purposes. One is to protect a domestic industry from foreign competition (a protective tariff), and the other is to raise revenue for the government (a revenue tariff). Whatever the tariff's purpose, the effects on the economy are the same.

The effects of a tariff are illustrated in Figure 14.6. Part (a) shows that under free trade, the country accepts the world price P_w which is lower than the domestic price P_d (HL students may note this means the country has a comparative disadvantage). At P_w it produces quantity Q_1, demands quantity Q_4, and imports $Q_4 - Q_1$. Suppose a tariff is imposed on the imported good. As a result, *the price of the imported good rises, to* $P_w + t$, causing the domestic price of the good to rise above the world price by the amount of the tariff, to $P_w + t$.

Effects of tariffs: winners and losers

At the new price $P_w + t$, domestic quantity supplied increases from Q_1 to Q_2, domestic quantity demanded falls to Q_3, and the quantity of imports falls to $Q_3 - Q_2$. We will now examine who gains and who loses from these changes.

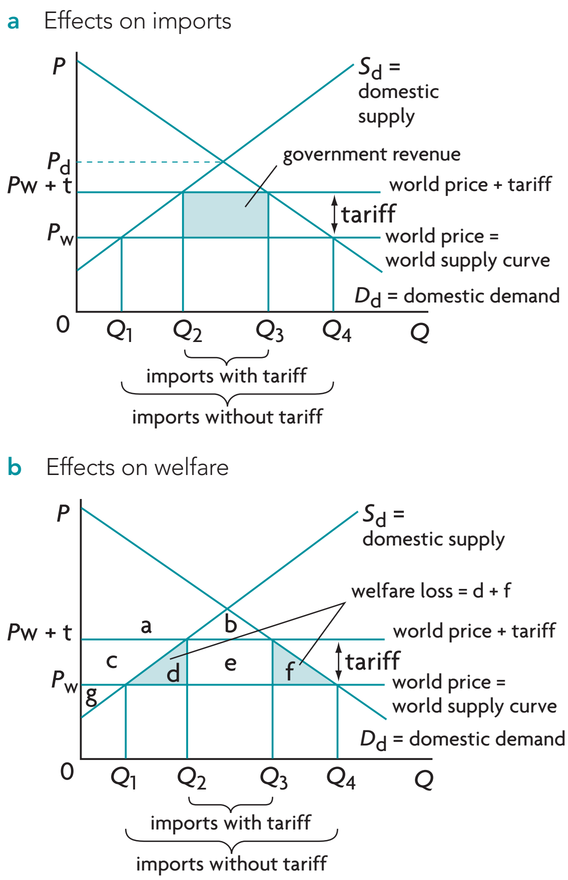


Figure 14.6: Effects of a tariff

Winners

- **Domestic producers are better off**. Domestic producers who receive the protection gain from the tariff, because they receive a higher price, P_w+t, and they sell a larger quantity, Q_2 (rather than Q_1).
- **Domestic employment in the protected industry increases**. Since domestic producers sell a larger quantity, this has the effect of increasing employment in the protected industry.
- **The government gains tariff revenues.** The amount of revenue the government receives from the tariff is shown by the shaded area in Figure 14.6(a), determined by multiplying the amount of tariff (per unit of the good) times the quantity of imports. Since the tariff is paid by consumers (who pay the price P_w+t), the government's tariff revenue represents income that is transferred from consumers to the government.

Losers

- **Domestic consumers are worse off**. Consumers lose from the tariff, because they must pay a higher price, P_w+t; and they can only buy a smaller quantity, Q_3 (rather than Q_4).
- **Domestic income distribution worsens**. There is a negative impact on income distribution, because the tariff is a type of regressive tax (see Chapter 12), which burdens people on lower incomes proportionately more than people on higher incomes; as income increases, the proportion of income paid as tax falls.
- **Increased inefficiency in production**. The increase in domestic output represents an increase in production by relatively inefficient domestic producers, resulting in a waste of scarce resources (inefficiency). Remember, the reason why the domestic price before trade is higher than the world price is that domestic producers are not as efficient as foreigners who produce and export the same good.
- **Foreign producers are worse off**. The producers of the exporting countries are worse off, because whereas they receive the world price, P_w, for their exports, they export a smaller quantity, since the quantity of imports in the importing country is reduced. The exporting countries therefore lose export revenues due to the fall in the quantity of exports.
- **A global misallocation of resource results**. The decrease in consumption, and the shift of production away from more efficient foreign producers and towards more inefficient domestic producers, indicate that there is an increase in the misallocation of resources both domestically and globally.

Welfare effects: the effects of tariffs on consumer and producer surplus

Figure 14.6(b) shows the effects of the tariff on consumer and producer surplus. Part (b) is identical to part (a), except that it labels the areas in the triangles and rectangles. Consumer surplus is the area under the demand curve and above the price paid by consumers (see Chapter 2). Before the imposition of the tariff it includes the areas a + b + c + d + e + f, representing the area under the demand curve and above the world price P_w. Producer surplus before the tariff is the area g below the price producers receive and above the supply curve. Therefore social (consumer plus producer) surplus is a + b + c + d + e + f + g.

After the tariff is imposed, consumer surplus drops to a + b, indicating that consumers are worse off and producer surplus becomes c + g, having increased by the amount of c, indicating that producers are better off. Also, the government gains the revenue equal to e. Therefore social surplus after the tariff is a+b+c+e+g.

To find the effect of the tariff on social surplus, we can subtract surplus after the tariff from surplus before the tariff to find the difference:

(a + b + c + d + e + f + g) − (a + b + c + e + g) = *d + f = welfare (deadweight) loss appearing as the shaded areas in the diagram.*

In effect consumers lost area c which was gained by producers, e which was gained by the government, as well as d due to inefficiency in production and f due to lower quantities. Therefore, the net loss is d + f. It results from a misallocation of resources caused by increased production by inefficient producers (area d) and decreased consumption of consumers (area f).

TEST YOUR UNDERSTANDING 14.6

1 Draw a diagram showing the effects of a tariff on
   - a the domestic price of the protected good,
   - b quantity produced domestically,
   - c quantity consumed by domestic consumers,
   - d quantity of imports, and
   - e government revenue from the tariff.

2 Using your diagram from question 1, discuss the effects of a tariff on
   - a domestic consumers,
   - b domestic producers,
   - c domestic employment,
   - d foreign producers,
   - e the government,
   - f efficiency in production, and
   - g the global allocation of resources.

3 In the United States, European meats, French chocolate and ham enjoy a 100% tariff while shelled peanuts enjoy a 131.8% tariff.[4] Evaluate the effects of these tariffs, taking into account various stakeholders, the domestic economy, the economies of exporters, and the global economy. Make sure you consider the welfare effects (effects on consumer surplus, producer surplus, and welfare loss).

Calculating the effects of tariffs from diagrams (HL only)

The tariff diagram in Figure 14.7 is similar to Figures 14.6, only we are given numerical data for prices and quantities of a good measured in millions of units. We would like to calculate the following information:

Tariff per unit

The tariff per unit is the world price + tariff minus the world price, or \$9 − \$7 = \$2 per unit.

Quantity of imports and import expenditures

Imports before the tariff are 2.8 million − 1 million = 1.8 million units. Imports after the tariff are 2.2 million − 1.5 million = 0.7 million units. Therefore imports fall by 1.8 million − 0.7 million = 1.1 million units.

Import expenditures before the tariff were P_W times the quantity of imports = \$7 × 1.8 = \$12.6 million but after the tariff they fell to P_W times the new quantity of imports = \$7 × 0.7 = \$4.9 million. (Note here that import expenditures after the tariff are calculated using

[4] United States International Trade Commission, referred to in https://www.businessinsider.com/americas-biggest-tariffs-2010-9?r=US&IR=T#french-jam-chocolate-and-ham-100-tariff-21

P_w, not $P_w + t$, because the t part of $P_w + t$ is the tariff per unit which is collected by the government.)

There was therefore a fall in import expenditures of $\$12.6 - \$4.9 = \$7.7$ million. (More simply the fall in import expenditures can be calculated as P_w times the fall in imports $= \$7 \times 1.1 = \$7.7$ million.)

Domestic consumers

The price paid by consumers increases from \$7 to \$9, or by \$2 per unit. The quantity purchased by consumers falls from 2.8 million units to 2.2 million units, or by 0.6 million units. Consumer expenditure before the tariff is $\$7 \times 2.8$ million $= \$19.6$ million, and after the tariff it is $\$9 \times 2.2$ million $= \$19.8$ million. Therefore consumer expenditure increases. Note however, that *this need not happen*. It is possible for consumer expenditure to fall, if the quantity effect on expenditure is larger than the price effect. *Consumers are worse off regardless, because they must pay a higher price for a smaller quantity*.

Domestic producers

The price received by producers increases from \$7 to \$9, or by \$2 per unit. The quantity produced also increases from 1 million to 1.5 million units, or by 0.5 million units. Therefore, producer revenue increases from $\$7 \times 1$ million units $= \$7$ million before the tariff to $\$9 \times 1.5$ million units $= \$13.5$ million after the tariff, or by \$6.5 million (= \$13.5 million − \$7 million). Producers are better off.

Government revenue

Government revenue from the tariff increases from zero to an amount equal to the tariff per unit (\$2) times the quantity of imports *after* the tariff has been imposed (0.7 million units). It is therefore $\$2 \times 0.7$ million $=$ \$1.4 million. The government budget therefore gains.

Consumer and producer surplus

Consumer surplus before the tariff is $(16.5 - 7) \times \frac{2.8}{2} =$ \$13.3 million. After the tariff it falls to $(16.5 - 9) \times \frac{2.2}{2} =$ \$8.25 million. Therefore consumer surplus falls by $13.3 - 8.25 = \$5.05$ million.

Producer surplus before the tariff is $(7 - 3) \times \frac{1}{2} =$ \$2 million, while after the tariff it increases to $(9 - 3) \times \frac{1.5}{5} = \$4.5$ million. Therefore producer surplus increases by $\$4.5 - \$2 = \$2.5$ million.

Welfare loss

Welfare loss can be calculated as the area of the two shaded triangles: $\left[\frac{(9-7)\times(1.5-1)}{2}\right] + \left[\frac{(9-7)\times(2.8-2.2)}{2}\right] = 0.5 + 0.6 = \$1.1$ million.

Foreign producers

Exports of foreign producers to the country that imposes the tariff fall by an amount equal to the fall in the country's imports, calculated above to be 1.1 million units. Export revenues of the foreign producers fall by an amount equal to the fall in the quantity of exports (1.1 million units) times the world price (\$7), which is \$7.7 million (= $\$7 \times 1.1$ million). Note that this is the same as the fall in import expenditures calculated above. Therefore foreign producers are worse off.

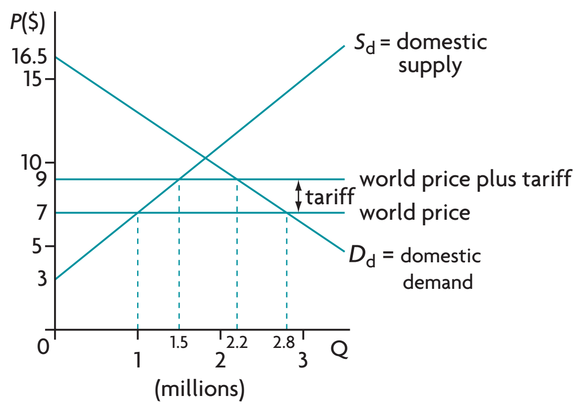


Figure 14.7: Calculating the effects of a tariff

TEST YOUR UNDERSTANDING 14.7

1 With no trade, the price of computers in Lakeland was \$400 per computer. Under free trade, the price of computers in Lakeland was \$300 per unit; domestic computer production was 100 000 units per year and imports stood at 250 000 units per year. Following the imposition by the government of a \$50 per unit tariff on computers, domestic production increased to 200 000 units and imports

CONTINUED

fell to 70 000 units. **i** Using the concept of comparative advantage and the prices of computers in Lakeland under no trade and under free trade, explain why Lakeland is an importer of computers. **ii** Calculate

- **a** using the new price paid by consumers,
- **b** calculate the new price received by domestic producers,
- **c** total domestic computer sales before the tariff, and
- **d** total domestic computer sales after the tariff.

2 Using your results in question 1, calculate the following effects of the imposition of the $50 per unit tariff on computers:

- **a** change in consumer expenditure,
- **b** change in import expenditure,
- **c** change in domestic producer revenue,
- **d** change in the government's revenue,
- **e** change in foreign producers' quantity of computer exports to Lakeland,
- **f** change in foreign producers' export revenues from computer exports to Lakeland,
- **g** change in consumer surplus,
- **h** change in producer surplus, and
- **i** welfare loss. (You may find it useful to use a tariff diagram to do your calculations. The diagram does not have to be drawn to scale.)

Import quotas

An **import quota** (or more simply, **quota**) is a legal limit to the quantity of a good that can be imported over a particular time period (typically a year). The effects of quotas are similar to the effects of tariffs, except that they usually do not create revenue for the government.

Figure 14.8(a) shows the effects of an import quota. Suppose initially the economy is importing under free trade; quantity Q_1 is supplied by domestic producers, quantity Q_4 is demanded, and the excess of quantity demanded over quantity supplied, $Q_4 - Q_1$, represents imports. (HL students may note that as in the case of tariffs, the country has a comparative disadvantage since P lies below the domestic price without trade.) The government then decides to impose a quota on imports, limiting the quantity that can be legally imported to $Q_3 - Q_2$. This means that a new supply curve is created by the addition of the amount of the quota to domestic supply. In other words, for each price, the quantity that is available to be purchased by domestic consumers is equal to the quantity produced by domestic producers (shown by the domestic supply curve S_d) plus the quantity of imports permitted by the quota. In Figure 14.8(a), the new, after-quota supply curve is shown by S_{dq}. S_{dq} begins at price P_w, indicating that it is not possible to import the good at a price lower than P_w (as no foreign producers would sell at a price lower than P_w). The new equilibrium domestic price is determined by the intersection of the domestic demand curve with S_{dq}, and is P_q.

When the government sets a quota, it issues a limited number of import licences determining the legal limit on the quantity of imports. The holders of these licences are the only individuals with the legal right to import. These licence holders gain quota revenues (also known as 'quotas rents') because whereas they buy the good at the world price P_w, they sell it to consumers at the higher domestic price P_q. Usually, the government gives the licences to governments of exporting countries, which then distribute them to their own producers or exporters, who buy at the price P_w and sell at P_q. As a result, the exporters (or producers) of exporting countries receive the quota revenues.[5] Because of this, foreign governments/producers prefer having quotas rather than tariffs imposed upon their exports.

Effects of a quota: winners and losers

With the exception of who gets the quota revenue, and welfare loss, the effects of an import quota are the same as in the case of a tariff.

[5] It is also possible for the government to sell the licences to domestic importers, in which case it gains the quota revenue; alternatively, the government could distribute the licences to domestic importers without charging a price, so the importers gain the quota revenues, because they buy the imports at price P_w and sell them at P_q. However, neither of these two options are commonly practised.

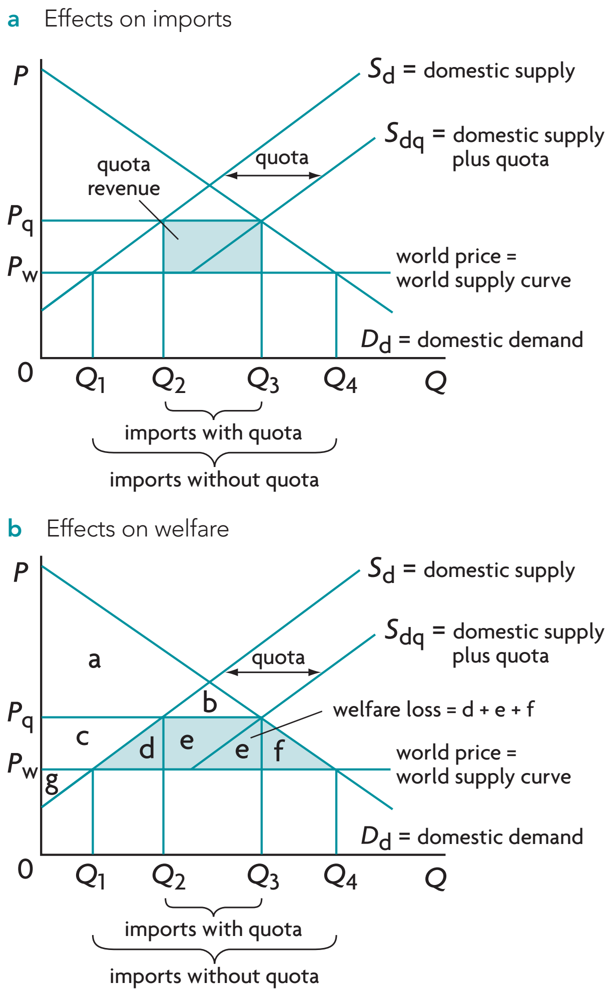


Figure 14.8: Effects of a quota

At the new supply curve, S_{dq}, domestic production increases to Q_2, domestic quantity demanded falls to Q_3, and the quantity of imports falls to $Q_3 - Q_2$.

Winners

- **Domestic producers are better off**. As in the case of a tariff, domestic producers who receive the protection gain from the quota, as they receive a higher price, P_q, and they sell a larger quantity, Q_2 (rather than Q_1).
- **Domestic employment increases.** As in the case of a tariff, domestic employment in the protected industry increases since producers increase the quantity of output they produce.

Neutral impact

- **The government neither gains nor loses.** Since the government usually gives the import licences to foreign governments, the government budget is not affected.

Losers

- **Domestic consumers are worse off.** As in the case of a tariff, consumers lose from the quota, because they must pay a higher price, P_q, and they can only buy a smaller quantity, Q_3 (rather than Q_4).
- **Domestic income distribution worsens.** Quotas do not involve a tax in the same way that tariffs do; however, they do result in a higher price, and the difference $P_q - P_w$, or the increase in price, has the same effect as the tariff in that it is regressive. In other words, the amount $P_q - P_w$ represents a higher fraction of income when income is low. Therefore quotas have the effect of worsening the distribution of income.
- **Increased inefficiency in production.** As in the case of tariffs, there results an increase in production by relatively inefficient domestic producers.
- **The exporting countries may be worse off or better off.** As in the case of a tariff, the producers of the exporting countries export a smaller quantity, resulting in a loss of export revenues. However, since the exporting countries receive the import licences, they gain the quota revenues. Therefore, whether they will be worse off or better off depends on which is larger: the loss of export revenues or the gain of quota revenues.
- **A global misallocation of resources results.** The decrease in consumption, and the shift of production away from more efficient foreign producers and towards more inefficient domestic producers indicates that there is an increase in the misallocation of resources globally, affecting both consumers and producers.

Welfare effects: the effects of quotas on consumer and producer surplus

The effects of quotas on consumer and producer surplus are shown in Figure 14.8(b), which is the same as part (a) apart from the labelling of the areas. The welfare effects of quotas differ from those of tariffs because of quota rents which are lost to the domestic economy (whereas tariff revenues come back to society due to government spending on merit goods). Before the quota, consumer surplus is the area a+b+c+d+e+f, and producer surplus by area g. After the quota, consumers have surplus equal to area a + b, and producers have surplus equal to g + c. Areas d and f have been lost as welfare loss, due to inefficiencies in production (area d) and reduced consumption (area f). Area e represents quota revenue that is transferred abroad to exporting countries. Therefore, the total surplus lost due to the quota is d + e + f. Therefore, quotas result in greater welfare losses for the domestic economy than tariffs.

A note on tariff and quota diagrams and their interpretation

While it is relatively straightforward to understand tariffs and their diagrams, quotas can sometimes be a little confusing. However, they also become very straightforward and simple to understand if one takes the following idea into consideration.

Tariffs and quotas are two different ways of achieving the same result, which is a *lower quantity of imports and a higher domestic price*. Tariffs work by increasing the price of imports, and then allowing demand and supply to arrive at the new, lower, quantity of imports. Quotas work by restricting the quantity of imports, and then allowing demand and supply to arrive at the new, higher, price of imports. This is why the effects of tariffs and quotas are the same (with the exception of quota rents and hence welfare loss).

This is also why the two diagrams are very similar. *The following simple instructions will show you how to convert a tariff diagram into a quota diagram*. Draw a tariff diagram as in Figure 14.6(a), which is reproduced by the green lines in Figure 14.9.

1 Delete the portion of the P_w+t curve that lies to the right of the D_d curve (the portion that is crossed out with red lines).
2 At the point where D_d intersects P_w+t, draw a line parallel to the S_d curve up to P_w and label it S_{dq}; this appears as a blue line in Figure 14.9.

You now have all the lines of the quota diagram that are needed. It is only necessary now to do some re-labelling.

3 Cross out P_w+t and re-label the price P_q.
4 Write in 'quota' between the two supply curves as in Figure 14.8.

You now have a quota diagram that is identical to the quota diagram in Figure 14.8.

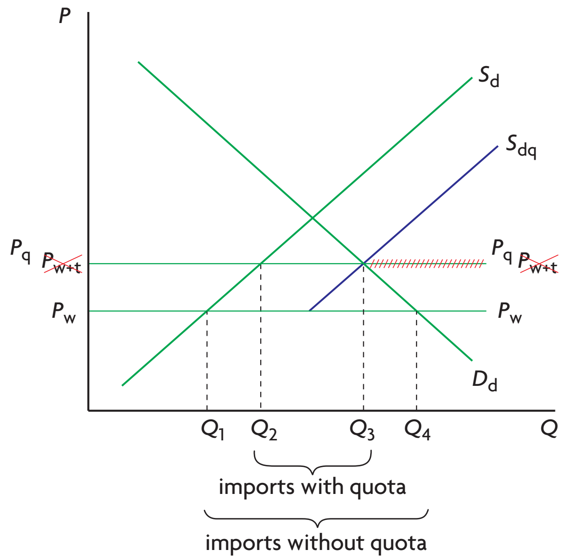


Figure 14.9: Converting a tariff diagram into a quota diagram

TEST YOUR UNDERSTANDING 14.8

1 Draw a diagram showing the effects of a quota on
   a the domestic price of the protected good,
   b quantity produced domestically,
   c quantity consumed by domestic consumers,
   d quantity of imports, and
   e quota revenue.
2 Using your diagram from question 1, and assuming the government gives quota licences to foreign firms, analyse the effects of a quota on
   a domestic consumers,
   b domestic producers,
   c domestic employment,
   d foreign producers (*Hint:* take into account loss of export revenues and gain of quota revenue),
   e the government,
   f efficiency in production, and
   g the global allocation of resources.

CONTINUED

3 a What is the main difference between the effects of a tariff and a quota?

b Explain why foreign producers are likely to prefer quotas to tariffs.

c Explain why tariffs are preferable from the point of view of the country that imposes trade protection measures.

4 The European Union imposes quotas on beef, according to which it distributes licences to certain exporting countries for specific beef quantities.[6] Evaluate the effects of such quotas taking into account various stakeholders, the domestic economy, the economies of exporters, and the global economy. Make sure you consider the welfare effects (effects on consumer surplus, producer surplus, and welfare loss).

Calculating the effects of quotas from diagrams (HL only)

The quota diagram in Figure 14.10 is similar to Figure 14.8, only we are given numerical data for prices and quantities of a good measured in millions of units. We would like to calculate the following information:

Import quota

The import quota can be read off as 16 million – 11 million = 5 million units, i.e. this is the permissible number of units that can enter the country per time period (such as a year).

The price after the quota is imposed

The price paid by consumers and received by *domestic* producers increases from €10 to €14.

Quantity of imports and import expenditures

Imports fall from = 15 million units (= 20 − 5) before the quota to 5 million units (= 16 − 11) after the quota (this is the number of units permitted by the quota). Therefore, imports fall by 10 million units.

Import expenditures before the quota were P_W times the quantity of imports = 10 × 15 = \$150 million and after the quota they were P_W times the new lower quantity of imports = 10 × 5 = \$50 million.

Therefore import expenditure fell by \$100 million (= 150 − 50). More simply the fall in import expenditures can be calculated as P_W times the fall in imports = 10 × 10 = \$100 million.)

Domestic consumers

The price paid by consumers increases from to €10 to €14, or by €4 per unit. The quantity purchased by consumers falls from 20 million units to 16 million units, or by 4 million units. Consumer expenditure before the quota is €10 × 20 million = €200 million, and after the quota it is €14 × 16 million = €224 million. Therefore, consumer expenditure increases. Note however, that as in the case of tariffs, *this need not happen*. Consumers are worse off because they must pay a higher price for a smaller quantity.

Domestic producers

The price received by producers also increases from €10 to €14, or by €4 per unit. The quantity produced increases from 5 million to 11 million units, or by 6 million units. Therefore, producer revenue increases from €10 × 5 million units = €50 million before the quota to €14 × 11 million units = €154 million after the quota, or by €104 million (= €154 million − €50 million). Producers are better off.

The government

The government budget is not affected.

Consumer and producer surplus

Consumer surplus before the quota was $(30 - 10) \times \frac{20}{2} =$ \$200 million, while after the quota it dropped to $(30 - 14) \times \frac{16}{2} =$ \$128 million. Therefore consumer surplus fell by \$200 − \$128 = \$72 million.

Producer surplus before the quota was $(10 - 7) \times \frac{5}{2}$ = \$7.5 million. After the quota it increased to $(14 - 7) \times \frac{11}{2}$ = \$38.5 million. Therefore producers gained surplus of \$38.5 − \$7.5 = \$31 million.

Welfare loss

Welfare loss in the case of quotas differs from the case of tariffs because of the quota rents that are transferred abroad to the exporting countries. Therefore welfare loss is equal to the areas d + e + f in Figure 14.8(b). Area

[6] http://beefandlamb.ahdb.org.uk/market-intelligence-news/eu-beef-quota-usage

d $= \frac{(14-10)\times(11-5)}{2}$ = \$12 million + area e = (14 − 10) × (16 − 11) = \$20 million + area f $= \frac{(14-10)\times(20-16)}{2}$ = \$8 million making a total welfare loss of \$40 million.

Areas d + e + f form a trapezium which can be more easily calculated by multiplying the average of the two parallel sides by the height. This gives $\frac{(5+15)}{2} \times 4$ = \$40 million. (Calculation of the area of a trapezium was explained in Chapter 2.)

Foreign producers

Exports of foreign producers to the country imposing the quota fall by an amount equal to the fall in the country's imports, calculated above to be 10 million units. Export revenues of the foreign producers fall by an amount equal to the fall in the quantity of exports (10 million units) times the world price (€10), which is €100 million (= €10 × 10 million). However, since the foreign producers receive the quota revenue (by receiving the quota licences), they gain €20 million (equal to the increase in price per unit due to the quota, or €4, times the number of units allowed by the quota, or 5 million). Therefore their losses are €80 million (= €100 million − €20 million). (It is possible for the gain from quota revenue to be greater than the loss of export revenues, so that foreign producers could be better off with the quota than without it.)

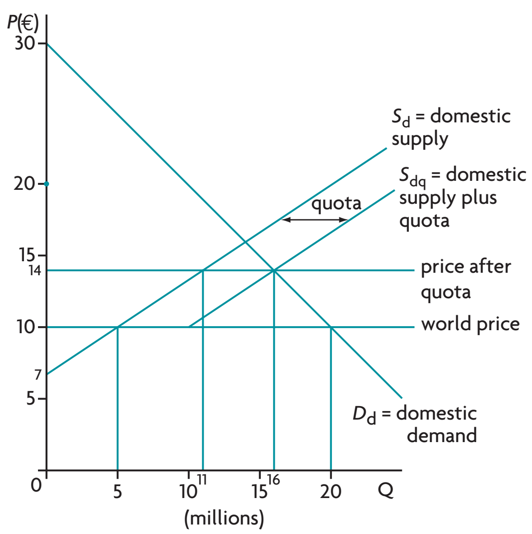


Figure 14.10: Calculating the effects of an import quota

TEST YOUR UNDERSTANDING 14.9

1 Under free trade, the price of mobile phones in Lineland is €100 per unit; domestic sales are 700 000 units per year, of which 500 000 units are imported. Following the imposition of an import quota of 200 000 units per year, domestic sales fall to 500 000 units per year, and the price increases to €120 per unit. Find the quantity of mobile phones produced domestically

   a before the quota, and

   b after the quota.

   c What is the quantity of imports after the quota?

   d Draw a quota diagram plotting these figures.

2 Using your results in question 5, calculate the

   a change in consumer expenditure,

   b change in import expenditure,

   c change in producer revenue,

   d quota revenue,

   e change in foreign producers' quantity of mobile phone exports to Lineland,

   f change in foreign producers' export revenues from mobile phone exports to Lineland,

   g change in consumer surplus,

   h change in producer surplus, and

   i welfare loss. (You may find it useful to use a quota diagram to do your calculations.)

3 Assuming that quota revenues are gained by foreign producers, explain

   a the effect on the government's budget, and

   b the total gains/losses of foreign producers, taking into account both export revenues and quota revenues.

REAL WORLD FOCUS 14.1

US Sugar Quotas

Introduction

As we will see later in this chapter, when trade is restricted there may be further negative effects that extend beyond the protected industry into the entire economy. One of these is increased costs of production that arise when a protected good is used as an input in production, where higher prices due to tariffs lead to cost-push inflation, associated job losses and loss of export competitiveness. In addition, protection by one country may lead to retaliation by other countries.

Figure 14.11: Jeanerette, Louisiana, USA. Sugarcane production factory

US sugar quotas

The sugar industry in the United States has been receiving protection since 1789. One of the features of protection is import quotas, which restrict the quantity of imports. The imports come from 40 countries, each of which is issued a quota licence specifying how much sugar they can export to the United States.

The result of the quantity restrictions is to raise the domestic price of sugar to about double the world price. This works to increase domestic sugar production, while more efficient sugar farmers in developing countries are deprived of export markets and export revenues. Of total US sugar consumption, 80% is produced domestically and 20% is imported.

The import quotas are supported by US sugar farmers, though they run against the interests of US consumers who must pay the higher price, as well as confectionery and soda makers who use sugar as an input.

The programme has negative overall employment effects. While jobs are created in the sugar-producing industry, many more are lost in the industries that rely on sugar. According to the American Enterprise Institute, about 10 000–20 000 jobs are lost each year. A number of candy factories have relocated to Mexico over the years as they found they were not profitable in the United States.

There are also environmental costs involved. Over half of sugar comes from sugar beets, grown on very fertile irrigated land that could have been used for other crops. The rest of the sugar comes from sugar cane that uses large amounts of nitrogen and other fertilisers, with serious effects on the water quality and natural ecology of the regions where they are grown.

Sources: *Jenny Grimberg, 'The Cost of Protecting the United States Sugar Industry' https://beinglibertarian.com; Vincent H. Smith, 'The U.S. Spends $4 billion a year subsidizing "Stalinist-style" domestic sugar production', Market Watch, 26 June, 2018, https://www.marketwatch.com/story/the-us-spends-4-billion-a-year-subsidizing-stalinist-style-domestic-sugar-production-2018-06-25*

Applying your skills

1 Using a quota diagram, explain why US sugar producers support the import quota.

2 Using an *AD-AS* diagram, explain why the quotas have negative overall effects on employment.

3 Explain what kind of unemployment has been created by the relocation of some confectionery factories to Mexico.

CONTINUED

4 Use an appropriate externality diagram to explain the environmental costs of sugar production.

5 Using a diagram, evaluate the effects of the sugar import quotas on various stakeholders, the US economy and sugar-exporting countries.

6 Evaluate the use of sugar import quotas in the United States.

Production subsidies

Subsidies were introduced in Chapter 4, where we saw that a subsidy is a payment by the government to a firm for each unit of output produced. In the context of trade protection there are two kinds of subsidies. One is intended to protect domestic firms that compete with imports, called a 'production subsidy', that we consider in this section; the other is a subsidy intended to protect domestic firms that export, called an 'export subsidy'. which we will consider in the next section.

In the context of trade protection, *production subsidies* are payments per unit of output granted by the government to domestic firms that compete with imports. In Figure 14.12 (a), under free trade, the country would produce quantity Q_1 of the good, quantity demanded would be Q_2, and excess demand of $Q_2 - Q_1$ would be satisfied by imports. Now suppose the government grants a subsidy to domestic firms per unit of output produced. We know from Chapter 4 that the subsidy causes the domestic supply curve to shift downward by the amount of the per unit subsidy, to S_{ds}. The good continues to sell domestically at the world price, P_w, though the price received by producers is now $Pw+s$.

Effects of production subsidies: winners and losers

The domestic firms supply the larger quantity Q_3, determined by the intersection of the after-subsidy supply curve S_{ds} with the world price line. As a result quantity of imports fall from $Q_2 - Q_1$ to $Q_2 - Q_3$.

Winners

- **Domestic producers are better off**. As a result of the subsidy, domestic producers in the protected industry receive the price P_w+s (= P_w plus the subsidy per unit), and domestic production expands from Q_1 to Q_3; therefore producers benefit.
- **Domestic employment increases**. The increase in domestic production from Q_1 to Q_3 causes domestic employment in the protected industry to increase.

Neutral impact

- **Consumers are not affected**. Consumption of the good both before and after the subsidy is at Q_2 units of output, and the price stays the same, at P_w. (Following the imposition of the subsidy, consumers buy more of the domestic good whose production has increased, and less of the imported good.)

Losers

- **The government budget**. The government budget is negatively affected as the government must spend tax revenues on the subsidy. The amount spent on the subsidy is $Pw+s-P_w$ (the subsidy per unit) times Q_3, or the quantity produced domestically.
- **Taxpayers are worse off**. Taxpayers lose as a portion of tax revenues is spent on production subsidies that have the effect of increasing production of inefficient producers. The amount lost is what is spent on the subsidy out of the government budget (see above). These funds could have been spent elsewhere with benefits for taxpayers (such as spending on merit goods).
- **Increased inefficiency in production**. As in the case of tariffs and quotas, production of domestic inefficient producers increases, while the production of more efficient foreign producers falls.
- **The exporting countries are worse off**. Foreign producers exporting the good are worse off because they can export less of the good, and export revenues of these countries fall.
- **A global misallocation of resources results**. The shift of production from efficient to inefficient producers involves an increase in the global misallocation of resources, negatively affecting economies.

It should be noted that the subsidies discussed here are granted on goods that are produced for the domestic market, as the entire quantity produced is sold domestically. The objective of such subsidies, as we have seen, is to raise the domestic price to producers and increase their quantity supplied, thus decreasing the quantity of imports, therefore providing protection to producers (and their workers).

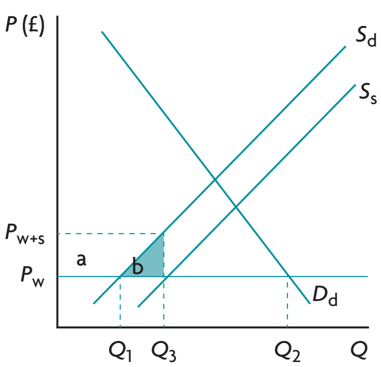
It is possible, however, if the subsidy is very large, for producers to increase their production by so much that quantity supplied becomes greater than all of domestic quantity demanded. In this case, the excess of quantity supplied over quantity demanded would be exported. A higher cost country becomes an exporter, with an even greater domestic and global misallocation of resources.

Welfare effects: the effects of production subsidies on consumer and producer surplus

Figure 14.12(a) also shows the effects of production subsidies on consumer and producer surplus. Consumer surplus is not affected, since both price paid and quantity bought by consumers have not changed. Producers, on the other hand, gain surplus given by the area a due to the higher price they receive and the larger quantity they sell. However, the government loses the area $(Pw+s-Pw)\times Q_3$ = a + b, due to government spending to provide the subsidy. Therefore by subtracting what is lost due to government spending from what is gained by producers, we find a net loss to society of the shaded area b.

From an efficiency point of view the effects of production subsidies may not be as harmful as those of tariffs and quotas, because while they encourage inefficient production (like tariffs and quotas) they do not have negative effects on consumption which remains the same before and after the subsidy. This can be seen from the loss of only producer surplus in Figure 14.12(a).

a Effects on imports and welfare



b Calculating the effects of a production subsidy

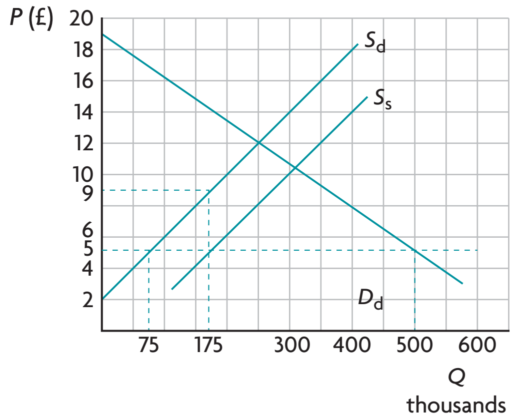


Figure 14.12: Effects of a production subsidy

TEST YOUR UNDERSTANDING 14.10

1 Draw a diagram showing the effects of a production subsidy on
   - a the domestic price of the subsidised good,
   - b quantity produced domestically,
   - c quantity consumed by domestic consumers, and
   - d quantity of imports.

2 Using your diagram from question 1, explain the effects of a production subsidy on
   - a domestic consumers,
   - b domestic producers,
   - c domestic employment,
   - d foreign producers,
   - e the government,
   - f efficiency in production, and
   - g the global allocation of resources.

3 The US government spends more than $20 billion a year on subsidies for agricultural products, many of which are exported in addition to being sold in the domestic market. Evaluate the effects of such production subsidies, taking into

CONTINUED

account various stakeholders, the domestic economy, the economies of exporters and the global economy.

4 Economists generally prefer production subsidies to tariffs and quotas. Using a subsidy diagram explain their reasoning.

Calculating the effects of production subsidies from diagrams (HL only)

The subsidy diagram in Figure 14.12(b) is similar to Figure 14.12(a), only we are given numerical data for prices and quantities of a good measured in millions of units. We want to calculate the following information:

Subsidy per unit

The subsidy per unit is the vertical difference between the two supply curves, or £9 − £5 = £4 per unit.

Quantity of imports and import expenditures

Imports before the subsidy are 425 thousand units (= 500 − 75). Imports after the subsidy are 325 thousand units (= 500 − 175). Therefore imports fall by 100 thousand units (= 425 − 325). Import expenditures fall by the decrease in imports (100 thousand units) times the world price (£5) = £500 thousand.

Domestic consumers

Consumers are not affected. They pay the same price, £5, and they buy the same quantity, 500 thousand units before and after the subsidy.

Domestic producers

The price received by producers increases from £5 to £9, or by £4 per unit, which is the subsidy per unit. The quantity produced also increases from 75 to 175 thousand units, or by 100 thousand units. Therefore producer revenue increases from £5 × 75 = £375 thousand before the subsidy to £9 × 175 = £1575 thousand after the subsidy, or by £1200 million (= £1575 − £375). Producers are better off.

Government expenditure and taxpayers

Government expenditure on the subsidy increases from zero to an amount equal to the subsidy per unit (£4) times the quantity produced domestically *after* the subsidy has been granted (175 thousand units). It is therefore £4 × 175 = £700 thousand. The government budget therefore loses. Taxpayers are worse off by the equivalent amount, as their tax funds are spent on subsiding inefficient producers with no benefits to themselves.

Consumer and producer surplus

Consumers buy the same quantity at the same price before and after the production, consumer surplus remains unchanged at $(19 - 5) \times \frac{500}{2}$ = £3500 thousand. Producer surplus initially before the production subsidy was $(5 - 2) \times \frac{75}{2}$ = £112.5 thousand. After the subsidy it is $(9 - 2) \times \frac{175}{2}$ = £612.5 thousand. Therefore producer surplus increases by 612.5 − 112.5 = £500 thousand.

Welfare loss

Welfare loss is the triangle given by $\frac{(9-5)\times(175-75)}{2}$ = £200 thousand.

Foreign producers

Exports of foreign producers to the country granting the subsidy fall by an amount equal to the fall in the country's imports (which is also equal to the increase in domestic production), calculated above to be 100 thousand units. Export revenues of the foreign producers fall by an amount equal to the fall in the quantity of imports (100 thousand units) times the world price (£5), which is £500 thousand. (Note that this is equal to the fall in import expenditures noted above.) Therefore foreign producers are worse off.

TEST YOUR UNDERSTANDING 14.11

1 With no trade, the price of televisions in Radioland was £300 per television. Under free trade, the price of televisions in Radioland was £200 per unit; domestic television production was 300 000 units and domestic television sales were 900 000 units. Following the imposition of a £50 per unit subsidy on televisions, domestic production increased to 550 000 units.

   i Using the concept of comparative advantage and the prices of televisions under no trade and under free trade, explain why Radioland is an importer of televisions.

CONTINUED

ii State/calculate the

a post-subsidy price paid by consumers,

b post-subsidy price received by domestic producers,

c pre-subsidy quantity of imports,

d post-subsidy quantity of imports, and

e post-subsidy television sales.

2 Using your results in question 1, state/calculate the

a change in consumer expenditures,

b change in import expenditures,

c change in producer revenue,

d change in the government's budget,

e change in foreign producers' quantity of television exports to Radioland,

f change in foreign producers' export revenues from television exports to Radioland,

g change in consumer surplus,

h change in producer surplus, and

i welfare loss. (You may find it useful to use a subsidy diagram to do your calculations. This does not have to be drawn to scale.)

Export subsidies

Export subsidies are similar to production subsidies in that they involve a payment by the government per unit of the subsidised good, except that now *the subsidy is paid for each unit of the good that is exported*. Export subsidies are illustrated in Figure 14.13, showing an exporting country that succeeds in increasing its volume of exports by granting an export subsidy.

We see that before trade the world price P_w is higher than the domestic price (given by the intersection of domestic demand D_d and domestic supply S_d), so we know that once the country opens itself up to trade, it becomes an exporter of the good. At the world price P_w domestic

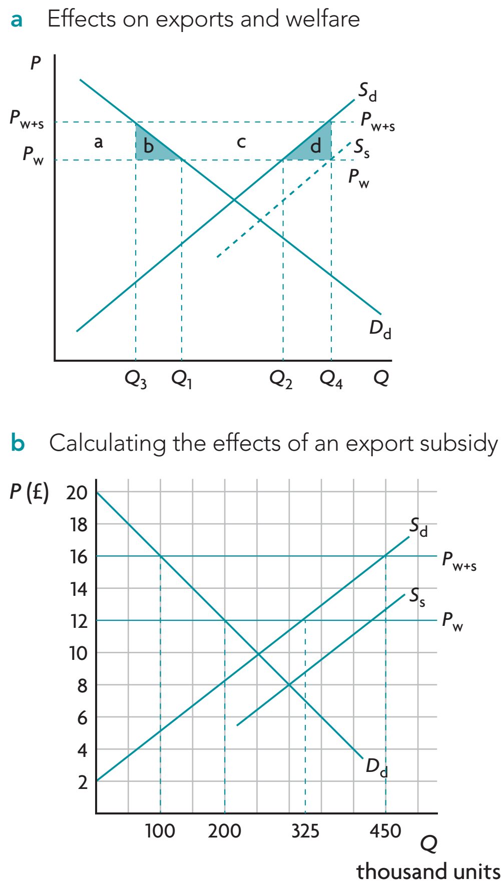


Figure 14.13: Effects of an export subsidy

quantity demanded is Q_1, domestic quantity supplied is Q_2, and $Q_2 - Q_1$ represents the quantity of exports.

The government now decides to grant an export subsidy per unit of the good exported in order to increase the quantity of exports. The supply curve shifts downward by the amount of the subsidy per unit to S_s. At the intersection of S_s with the world price line P_w, we draw a vertical line upward to find the new higher domestic price, P_{w+s}, which is equal to the world price P_w plus the subsidy per unit. At the higher price P_{w+s}, producers increase the quantity they supply to Q_4, while consumers decrease the quantity they demand to Q_3. The quantity $Q_4 - Q_3$ now represents exports. The price paid by foreigners remains at the world price, P_w. In other words, *the higher domestic price P_{w+s} applies only to domestic producers and consumers*.

(You may notice that with an export subsidy, consumers pay the higher price of P_{w+s}, in contrast to production subsidies where the price paid by consumer remains unchanged at P_w. This is because the export subsidy has the effect of reducing the quantity of the good available in the domestic market from Q_1 to Q_3, thus resulting in an increase in price (an upward movement along the demand curve). By contrast, with a production subsidy granted in a country that is importing the good, the increase in the quantity of imports satisfies the full amount of domestic quantity demanded at the world price P_w, therefore there is nothing to push the price up.)

Effects of export subsidies: winners and losers

We now examine who gains and who loses from the export subsidy.

Winners

- **Producers are better off**. Producers receive a higher price, P_{w+s}, and sell a larger quantity, Q_4 rather than Q_2, since the volume of exports increases from $Q_2 - Q_1$ to $Q_4 - Q_3$.
- **Domestic employment increases**. The increase in domestic production causes domestic employment to increase.

Losers

- **Consumers are worse off**. Consumers must pay a higher price for the good, P_{w+s}, and they consume a smaller quantity, Q_3 rather than Q_1.
- **Negative effect on the government budget.** The government must pay for the subsidy an amount which is equal to the subsidy per unit times the quantity of exports.
- **Taxpayers are worse off**. Taxpayers must pay indirectly for the subsidy, as the subsidy is financed out of tax revenues; moreover they lose by not having the subsidy funds available for alternative uses (such as merit goods).
- **Domestic income distribution worsens**. The reason for the worsening of income distribution is the same as in the case of quotas. While there is no regressive tax (as in the case of tariffs), consumers do have to pay a higher price due to the export subsidy, and the increase in price, represented by $P_{w+s} - P_w$, is regressive because it is a higher fraction of lower incomes than of higher incomes.
- **Increased inefficiency in production**. As in the case of all trade protection considered (tariffs, quotas, production subsidies) inefficient domestic producers are protected by the higher price.
- **The exporting countries are worse off**. Foreign producers are worse off as they lose a share of their global market through the increase in subsidised exports, and their export revenues fall.
- **There is an increase in the global misallocation of resources**. Consumers and producers around the world are negatively affected since the inefficiency of resource allocation around the world increases. In fact, the negative effects of export subsidies can be very serious. Real world focus 14.2 provides more details about these effects.

Welfare effects: the effects of export subsidies on consumer and producer surplus

Figure 14.13(a) shows the welfare effects of the export subsidy. Consumers lose areas a + b due to the higher price they must pay and the lower quantity they buy. Producers gain areas a + b + c as they are now receiving a higher price and producing a larger quantity. The government loses areas b + c + d, which is the amount they must pay for the subsidy, equal to the subsidy per unit times the quantity of exports, $Q_4 - Q_3$. Therefore the net loss to society is equal to the gain of producers minus the losses of consumers plus the government = (a + b + c) − (a + b) − (b + c + d) = − (b + d). Welfare loss to society consists of the two shaded triangles, b and d.

Note that the welfare loss from export subsidies are greater than those of production subsidies, because with export subsidies not only producers but also consumers are worse off.

TEST YOUR UNDERSTANDING 14.12

1. Draw a diagram showing the effects of an export subsidy on
   a. the domestic price of the subsidised good,
   b. quantity produced domestically,
   c. quantity consumed by domestic consumers, and
   d. quantity of exports.
2. Using your diagram from question 1, discuss the effects of an export subsidy on
   a. domestic consumers,
   b. domestic producers,
   c. domestic employment,
   d. foreign producers,

CONTINUED

e the government,

f efficiency in production, and

g the global allocation of resources.

3 India provides export subsidies for a number of products including pharmaceuticals, chemicals, information technology products, textiles and apparel. According to the WTO only countries with GNI *per capita* less than \$1000 can use export subsidies, and India does not qualify.[7] Evaluate the effects of such export subsidies, taking into account various stakeholders, the domestic economy, the economies of exporters and the global economy.

Calculating the effects of export subsidies from diagrams (HL only)

The subsidy diagram in Figure 14.13(b) is similar to Figure 14.13(a), only we are given numerical data for prices and quantities of a good measured in millions of units. We want to calculate the following information:

Subsidy per unit

The subsidy per unit is the vertical difference between the two supply curves, or £16 − £12 = £4 per unit.

Quantity of exports

Exports before the subsidy are 125 thousand units (= 325 − 200). Exports after the subsidy are 350 thousand units (= 450 − 100). Therefore, exports increase by 225 thousand units (= 350 − 125). Export revenues increase by the increase in exports (225 thousand units) times the world price (£12) = £2700 thousand.

Domestic consumers

Consumers pay a higher price, £16 rather than £12 and they buy a smaller quantity, 100 thousand rather than 200 thousand units. Consumer expenditure before the subsidy was 12 × 200 = £2400 thousand whereas after the subsidy it is 16 × 100 = £1600. (As noted above whether consumer expenditure increases or decreases depends on whether the price or the quantity decrease is larger.)

Domestic producers

The price received by producers increases from £12 to £16, or by £4 per unit, which is the subsidy per unit. The quantity produced also increases from 325 to 450 thousand units, or by 125 thousand units. Therefore producer revenue increases from £12 × 325 = £3900 thousand before the subsidy to £16 × 450 = £7200 thousand after the subsidy, or by £3300 thousand (= £7200 − £3900). Producers are better off.

Government expenditure and taxpayers

Government expenditure on the subsidy increases from zero to an amount equal to the subsidy per unit (£4) times the quantity of exports (350 thousand units). It is therefore 4 × 350 = £1400 thousand. The government budget therefore loses. Taxpayers are worse off by the equivalent amount, as their tax funds are spent on subsiding inefficient producers with no benefits to themselves.

Consumer and producer surplus

Consumer surplus falls from $(20-12)\times\frac{200}{2}$ = £800 to $(20-16)\times\frac{100}{2}$ = £200 thousand due to the export subsidy.

Producer surplus initially before the export subsidy was $(12-2)\times\frac{325}{2}$ = £1625 thousand. After the export subsidy it is $(16-2)\times\frac{450}{2}$ = £3150 thousand. Therefore, producer surplus increases by 3150 − 1625 = £1525 thousand.

Welfare loss

Welfare loss is given by the two shaded triangles given by

$$\left[\frac{(16-12)\times(200-100)}{2}\right]+\left[\frac{(16-12)\times(450-325)}{2}\right]$$

= 200 + 250 = £450 thousand.

[7] https://www.thehindubusinessline.com/economy/us-challenges-indias-export-subsidy-program-atwto/article23247487.ece; https://www.thehindubusinessline.com/opinion/export-subsidies-shouldgo/article23349292.ece

TEST YOUR UNDERSTANDING 14.13

1 With no trade, the price of fish in Oceanland was $2 per kilogram (kg). Under free trade, the price of fish was $3 per kg, Oceanland's fish production was 700 000 kg and fish consumption was 400 000 kg. After the government granted an export subsidy of $1 per kg, Oceanland's fish production increased to 900 000 kg while fish consumption fell to 250 000 kg.

   i Using the concept of comparative advantage and the prices of fish under no trade and under free trade, explain why Oceanland is an exporter of fish.

   ii Calculate

      a the post-subsidy price paid by domestic consumers,

      b post-subsidy price received by domestic producers,

      c pre-subsidy quantity of exports, and

      d post-subsidy quantity of exports.

2 Using the information and your results in question 1, state/calculate the

   a change in consumer expenditures,

   b change in producer revenue,

   c change in export revenue,

   d change in the government's budget,

   e change in consumer surplus,

   f change in producer surplus, and

   g welfare loss. (You may find it useful to use a subsidy diagram to do your calculations. This does not have to be drawn to scale.)

REAL WORLD FOCUS 14.2

Export subsidies: A WTO success story

Note: The World Trade Organization (WTO) is an international organisation responsible for overseeing the international trade system. We will examine the WTO in Chapter 15.

Export subsidies imposed by developed countries on their agricultural products have been popular both in developed but also in developing countries. For developed countries, the reason is obvious: farmers clearly benefit as they receive a higher price for the product, and production increases as the quantity of exports goes up. To understand the reasoning from the point of view of developing countries, we must go a little beyond our model of export subsidies of Figure 14.13.

Figure 14.14: Geneva, Switzerland. World Trade Organization (WTO) headquarters

In our analysis of trade protection, we have made the simplifying assumption that the importing or exporting country is so small that it cannot make an impact on the world price. However, in the real world we often deal with countries such as the United States, or groups of countries such as the European Union, which are very large. Their share of trade is sometimes so large that they do have an impact on the world price. In the case of US and EU exports, the quantities exported are sometimes so large that they often have the effect of lowering the world price. Therefore, export subsidies imposed by large countries, by increasing exports, have the effect of further lowering the world price, and hence the price of imports in developing countries.

CONTINUED

These lower import prices have very different effects on consumers and producers in developing countries. Consumers are better off, because they can buy food more cheaply. But producers are hurt because they are forced to compete with lower priced imports. This lowers their incentive to produce, so production falls, it lowers their revenues, it lowers their standard of living, and in many cases forces them to go out of business. To make matters worse, this leads to the country's growing reliance on food imports, which is hardly desirable in poor countries that sometimes suffer from chronic food shortages.

In view of this, why did developing country governments like developed country export subsidies? It is because they wanted cheap imported food to keep consumers happy. Food is an important political issue in many poor countries and high food prices can cause social and political unrest.

There is an irony in that rich country policy-makers cared more about farmers who are a tiny fraction of the population, accounting for about 2–4% of employment in the United States and European Union, while developing country policy-makers cared less about farmers who comprise as much as 40–60% (or more) of total employment.

We can see clearly that export subsidies have highly damaging effects because they:

- reduce competitiveness of developed country farmers through artificially low prices
- allow inefficient rich country farmers to gain market shares in the agricultural sectors of poor countries
- reduce standards of living of poor country farmers
- reduce food production and food security in poor countries while increasing reliance on imports.

At a WHO conference held in Nairobi in December 2015, an agreement was reached to eliminate export subsidies. This was hailed as a great success and a long overdue outcome for WHO negotiations. Whereas the GATT, the precursor of the WTO, had prohibited export subsidies (as well as import quotas) on manufactured products, it had permitted them for agricultural products due to pressures from developed countries responding to the special interests of their farmers.

At the same time that developed countries have phased out their export subsidies, these have begun to be applied by a number of developing countries, for example China for cotton, India for sugar, and Vietnam and Thailand for rice. Some developing countries object to the WTO prohibition of export subsidies, arguing that developing countries should be excluded from the prohibition.

Sources: *Heinz Strubenhoff, 'The WTO's decision to end agricultural export subsidies is good news for farmers and consumers', Brookings, www.brookings.edu/blog/future-development/2016/02/08/ Kimberly Ann Elliot, 'The WTO, Agriculture and Development: A Lost Cause?' International Centre for Trade and Sustainable Development, Bridges Africa, Volume 7, Number 1, www.ictsd.org/bridges-news/*

Applying your skills

1. Using demand and supply analysis, explain the effect that increases of exports by large countries can have on the world price of a good.
2. Using the information in the text, explain why export subsidies are potentially more harmful than production subsidies.
3. Using an export subsidy diagram discuss the impacts of an export subsidy on major stakeholders.

CONTINUED

4 Why do you think rich country policy-makers care so much about their farmers even though they are such a small fraction of the working population?

5 Evaluate the possible consequences of the WTO agreement to end export subsidies.

6 Some developing countries argue that the prohibition of export subsidies should not apply to them on the grounds that as they are making efforts to grow and develop, they should be entitled to benefits that have been enjoyed by developed countries for decades. Discuss this point of view.

Administrative barriers

Whenever a good is imported from another country, it must go through a number of customs procedures involving inspections, valuation (determining the value of the good), and others. In an effort to impose obstacles to imports and reduce their quantity, countries may increase the amount of red-tape checks and procedures, making them very time-consuming and difficult. In addition, importing countries can impose requirements that imported goods must be packaged in particular ways. Since exporters do not always fulfil the requirements, the quantity of imports is reduced.

In addition, many countries impose requirements that imported goods must fulfil particular technical standards, which involve health, safety and environmental conditions. In many cases, these standards automatically eliminate a range of imports. In other cases, certain products must undergo testing and inspection procedures that are so costly and time-consuming that once again the effect is to reduce the quantity of imports.

All the above procedures and requirements are known as **administrative barriers**.

In some cases, the imposition of such standards is justified by governments' concern for the health and safety of the domestic population, as well as possible negative environmental effects of imported goods. However, it is generally believed that the excessive use of these kinds of measures by governments is a disguised attempt to limit imports, and therefore is a kind of 'hidden' trade protection. See Real world focus 15.1 (Chapter 15) for an example.

Administrative barriers, like other forms of trade protection, have the effects of increasing the domestic price of the imported good, protecting inefficient producers and increasing allocative inefficiency.

TEST YOUR UNDERSTANDING 14.14

Research and find real world examples of administrative import procedures and health, safety and environmental standards applied to imported goods that are likely to be a form of hidden trade protection. Using your findings, explain their likely effects on stakeholders in the domestic and global economies.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 What are some of the most important exports and imports, as a share of total exports and imports, in the country you live in? Can you use the theory of comparative advantage to explain some of the patterns of trade (exports and imports) that you observe?

2 What countries are the most important trading partners of the country you live in? Can you identify some benefits that arise from trade with these countries?

CONTINUED

3 Research real-world examples of countries that take advantage of their factor endowments. Identify the endowments and explain how they are related to exports.

4 Select some of the more important imports in your country, and investigate what, if any, types of trade protection measures these are subject to. Identify at least one product that is subject to a tariff, a quota and a subsidy. Examine the likely reasons that this protection has been offered. Try to explain the effects these measures are likely to have on the various stakeholders and the economy you are studying.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 15

International trade: Part II

BEFORE YOU START

- How valid do you think are arguments in favour of trade protection (trade restrictions)?
- When countries engage in reducing or increasing trade barriers, what stakeholders do you think are affected and how might they become better off or worse off?
- What do you think is the meaning of economic integration?

This chapter continues the discussion of Chapter 14. We will review the arguments in favour for and against trade protection. We will then consider various forms of economic integration, that involve the removal of trade barriers and the promotion of free trade, either among small country groupings or the global economy through agreements involving many countries around the world. We will also consider monetary union, where countries give up their national currency by adopting a single currency with a unified monetary policy.

Photo: Sao Paolo, Brazil. Freighter loaded with containers

15.1 Arguments for and against trade protection

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and provide examples of arguments in favour of trade protection (AO2)
- explain and provide examples of arguments against trade protection (AO2)
- evaluate free trade versus trade protection, referring to the following: (AO3)
  - the arguments in favour of and against trade protection discussed in this section
  - the benefits of trade discussed in Chapter 14, Section 14.1
  - the advantages and disadvantages of the various trade protection measures discussed in Chapter 14, Section 14.3

Arguments for trade protection

In Chapter 14 we examined a variety of measures that governments use to create barriers to international trade. These barriers create some winners and some losers, but in all cases result in inefficiency in production and global resource misallocation. Why, then, do governments around the world continue to use trade protectionist policies?

Arguments that economists justify under certain conditions

These are arguments that may have validity under certain conditions. Their validity may depend on non-economic considerations, or on the expectation that longer-term economic benefits of trade protection are greater than short-term economic costs.

Infant industry argument

An **infant industry** is a new domestic industry that has not had time to establish itself and achieve efficiencies in production, and may therefore be unable to compete with more 'mature' competitor firms from abroad. Mature foreign firms, operating with lower costs of production, are able to sell at lower prices; domestic firms, being unable to compete, are unable to grow and may be forced to shut down. This argument rests on the principle of economies of scale, according to which a firm achieves lower average costs as it grows in size and produces more output. Therefore, a new firm with high costs of production that has not yet grown in size may need protection from imports until it grows to a size where protection is no longer needed.

This argument was first used in 1791 by Alexander Hamilton, the first US Secretary of the Treasury, to introduce tariffs to protect US industry and promote economic growth. Today it is used mainly for developing countries trying to expand their production into new areas and industries. Economists consider it to be one of the strongest arguments in favour of trade protection with a theoretical justification. It is justified on the grounds that a country may have a comparative advantage in the production of a particular *industrial* good, but cannot specialise in it unless it first receives some protection. However, the protection offered to infant industries must be temporary. Over a longer period, once the industry matures, the protection must be eliminated and the industry must compete in global markets under conditions of free trade.

In spite of its strong theoretical justification, there are some dangers in the infant industry argument. One is that it may be difficult for governments to know which particular industries have the potential to become low cost producers. Another is that once the selection of an industry is made, industries protected from competition may not have a strong incentive to become efficient. A third is that governments may continue to protect an industry even long after it has matured and is no longer an infant.

National security

According to this argument, certain industries are essential for national defence (such as aircraft, weapons, chemicals, certain minerals), and should be protected so that a country can produce them itself. In times of war or a national emergency, a country should not have to depend on imports for its defence. Moreover, there may be dangers in having 'unfriendly' nations specialise in weapons production.

While there is some merit to this argument, a problem is that it can be used by industries that have an indirect use in defence (such as the steel industry) to try to acquire protection against foreign competition. In 2018, the United States imposed highly controversial tariffs on steel and aluminium imports from a number of countries, which ended up provoking a trade war (see Real world focus 15.2). These tariffs were justified by reference to national security, though it is likely that other factors were at play. In the past, goods like candles, gloves, umbrellas, plastics, and others have received protection on the grounds that they were needed for national defence.

The national defence argument is a non-economic one, and so decisions should be made on political and military, not economic, grounds. Yet it is sometimes difficult to draw the line between what is essential for national defence and what is not.

Health, safety and environmental standards

Many countries maintain health, safety and environmental standards that all imported products must meet before they are allowed to enter. Each country sets its own standards, and governments are justifiably concerned that imported goods may fall short of these. However, there is a concern that these standards may sometimes be used as a form of 'hidden' protection to keep certain goods out if they are competing with domestically produced goods, such as administrative barriers (see Chapter 14).

Efforts of a developing country to diversify

Diversification means change involving greater variety; economic diversification refers to increasing the variety of goods and services produced; it is the opposite of *specialisation*. (HL students may note that diversification may not be consistent with the theory of comparative advantage.) A number of developing countries, especially **economically least developed countries** *(ELDCs)* which are among the poorest countries in the world, are very highly specialised in producing and exporting only a few primary commodities (for examples see Table 19.1, Chapter 19). Such excessive specialisation carries with it dangers, and countries may be better off diversifying their production and exports (to be discussed in Chapters 19 and 20). To be able to diversify, countries may have to use trade protection policies to keep out imports of goods they would like to produce themselves. For example, if a country would like to diversify into production of computers, it will have to impose barriers on imports of computers; alternatively the government could provide subsidies to domestic computer producers.

This argument applies only to developing countries. However, there may be a risk that governments may not know which products or industries are the most appropriate to select for protection that will allow for successful diversification.

Arguments that economists find questionable

Questionable arguments have limited validity, though they may have some value under special circumstances in offering short-term, temporary solutions to problems.

Anti-dumping

Dumping refers to the practice of selling a good in international markets at a price that is below the cost of producing it (usually by providing export subsidies, see Chapter 14). Dumping is considered to be an unfair trade practice, and is illegal according to international agreements. Nonetheless, it is a practice that continues to be used. According to the anti-dumping argument in favour of trade protection, if a country suspects that a trading partner is practising dumping, it should have the right to impose tariffs or quotas in order to limit imports of the subsidised, or dumped good; this is the **anti-dumping** argument in favour of trade protection.

The main problem with this argument is that because of difficulties involved in proving that dumping is being practised, many governments often use it as an excuse to offer protection to their domestic producers when this protection is not necessary or justifiable.

Unfair competition

Unfair competition refers to practices that countries may use in order to gain a competitive advantage over other countries in order to unfairly increase their exports at the expense of other countries. Examples include:

- dumping, discussed above
- the more general use of production and export subsidies whereby exporting firms artificially achieve lower costs of production thus increasing their exports (see Chapter 14)

- administrative barriers or 'hidden protection' whereby countries limit their imports using questionable means (see Chapter 14)
- undervalued currencies whereby countries seek a lower value for their currency in order to make their exports more competitive in foreign markets (see Chapter 16)
- violation of intellectual property whereby firms or individuals within countries illegally obtain ideas, trade secrets, inventions or anything else which is a creation of the human mind (intellectual works) and then use these to their own advantage.

The problem here is similar to the case of dumping, which is difficult and time consuming to prove. As a result, trading partners may take advantage of such arguments to impose protectionist measures, using them as an excuse to protect their own industries when protection is not justifiable.

Correcting a balance of payments deficit

A balance of payments deficit occurs when the outflow of money from a country is greater than the inflow, and usually happens when there are more imports than exports (see Chapter 16). It would seem that a way to correct this problem would be to impose barriers to the entry of imports into the country, limiting imports and therefore the need to make payments abroad. However, decreased imports would come at the expense of falling exports in exporting countries, and there is a risk of retaliation. Trade protection could be used as a short-term emergency measure if there is a serious balance of payments deficit; over the longer term there are other, more effective ways to deal with this problem (see Chapters 16 and 17).

Tariffs as a source of government revenue

Tariffs as a source of government revenue were common in the early stages of development of currently more developed countries. In the United States, for example, tariff revenues accounted for 56% of federal (central) government revenues in 1880; by 1900 these were 41%, and by 2000 they had fallen to less than 1%. Today, the use of tariffs for revenue purposes is more frequent in developing countries, where tariff revenues can sometimes account for as much as half or more of all government revenues. The reason for strong reliance on tariffs for revenues is related to the ease with which imports can be taxed, since they are goods that must pass through borders where they can be monitored.[1]

However, tariffs have disadvantages, as they are a regressive type of tax, and so have negative impacts on income distribution; in addition, they have negative effects on allocative efficiency. The convenience of relying on tariff revenues may also work as an excuse for governments to delay tax system reform. Therefore reliance on tariffs as a source of government revenues should be a temporary measure to be gradually phased out as countries grow and develop.

Protection of domestic jobs

According to this argument, restrictions on imports are needed to protect domestic employment. Import restrictions increase domestic production, thus increasing employment.

One problem with this argument is that if import restrictions apply to goods that are used as inputs in the production of other goods, it is likely that there will be higher costs of production due to the higher prices of the imported inputs, lower production in these industries, and therefore increased unemployment. It is even possible for the negative employment effects in the broader economy to be greater than the positive employment effects in the protected industry. See Real world focus 14.1, Chapter 14, for an example.

Another problem with this argument is that if unemployment in the domestic economy falls due to import restrictions, this means that unemployment increases in the countries that are forced to export less. The foreign countries that are hurt may retaliate by imposing import restrictions of their own. If the government's objective is to increase employment in the economy, fiscal, monetary or supply-side policies may be more appropriate. If, on the other hand, the government wants to increase employment in a particular industry, a subsidy is likely to be more appropriate than import restrictions (tariffs and quotas), because subsidies have fewer negative effects (see Chapter 14).

[1] By contrast, income taxes, which currently make up the largest share of government revenues in developed countries, are more difficult to levy and collect in less developed countries, partly because large shares of the population survive on very low incomes, partly because a very large proportion of the population are self-employed and working in the informal sector where taxes are not collected, and partly because of poor enforcement of tax collection and high tax evasion rates.

REAL WORLD FOCUS 15.1

Australia wages war on hidden protection

There is a concern in Australia that the use of hidden barriers in a number of its trading partners is growing. This includes even some countries with which Australia has free trade agreements.

According to the Australian Trade Minister, 'Tariffs and quotas are there for the whole world to see, but as we continue to shift towards a more open global economy, we're seeing a trend towards the use of hidden or invisible trade barriers.'

According to an analysis conducted by *Meat and Livestock Australia* and the *Australian Meat Industry Council*, non-tariff barriers in 41 countries are costing $3.4 billion a year.

Figure 15.1: Proserpine, Queensland, Australia. Loading sugarcane

The Australian Trade Minister said, 'We absolutely recognise the right of countries to protect consumers and the environment, through packaging requirements, shipment inspections, health and safety controls . . . Whilst many of these requirements are legitimate, we're seeing a shift towards how regulations and red tape are being manipulated and used as a deliberate tool to distort trade flows and protect domestic industries.'

The Australian government plans to use diplomatic pressure on the countries suspected of using hidden trade protection measures.

Source: *Brad Thompson 'Australia starts war on the rise of hidden trade barriers, plans diplomatic pressure', Financial Review, 7 December 2018. www.afr.com*

Applying your skills

1 Using your knowledge of trading blocs, (after you have read Section 15.2 below) explain why countries with which Australia has free trade agreements might face even greater incentives to use hidden trade barriers.

2 Outline why hidden trade protection is a type of unfair competition.

3 What did Australia's Trade Minister mean with his statement that hidden protection measures 'distort trade flows and protect domestic industries'?

Arguments against trade protection

Economists are most often critical of trade protection because it worsens the allocation of resources and imposes a variety of costs on the domestic and the global economies. In Chapter 14 we saw in some detail who gains and who loses from the imposition of a variety of protectionist measures in international trade. Table 15.1 provides a summary of the effects.

Impact on stakeholders	Tariffs	Quotas	Production subsidies	Export subsidies	Administrative barriers
Producers	gain	gain	gain	gain	gain
Workers	gain	gain	gain	gain	gain
Government	gain	neutral	lose	lose	neutral
Taxpayers	neutral	neutral	lose	lose	neutral
Consumers	lose	lose	neutral	lose	lose
Domestic society Producer efficiency	lose	lose	lose	lose	lose
Domestic society Income distribution	lose	lose	neutral	lose	lose
Domestic society Resource allocation	lose	lose	lose	lose	lose
Foreign producers	lose	lose	lose	lose	lose
Global society Resource allocation	lose	lose	lose	lose	lose

Table 15.1: Summary: who wins and who loses from trade barriers

Based on the information in Table 15.1, we may note the following:

- **Producers and workers (through the increase in domestic employment) are the only stakeholders who gain from all types of trade protection.** This is hardly surprising, since protectionist policies are usually undertaken with a view to protecting domestic production and domestic employment.
- **However, the gain of producers has a cost in terms of higher costs of production and reduced efficiency.** These result from all types of trade protection. In fact, long-term reliance on protection against foreign competition eliminates incentives for firms to lower costs and operate efficiently.
- **Consumers lose in most cases.** This is due to higher prices of protected goods and lower quantities of goods available in the market (the only exception is production subsidies, where the quantity consumed and price paid by consumers remain unaffected). Consumers also face reduced consumer choice through lower imports. The losses experienced by consumers as a rule are greater than the benefits to producers, confirmed by many studies that measure the effects of trade restrictions.
- **Income distribution in most cases worsens.** The only exception is production subsidies, where the price paid by consumers does not change.
- **Foreign producers are worse off in all cases.** This is apparent from the reduction in imports of the country imposing protectionist measures.
- **Domestic and global resource allocation lose under all forms of trade protection.** Domestically, the appearance of welfare loss means resource misallocation. In addition, there is resource misallocation on a global scale as protection means that production moves away from lower cost producers in other countries to higher cost producers who are enjoying the protective measures.

There are some additional points worth mentioning:

- **Looking at the broader economy, beyond the protected industry, trade protection may have negative effects on the price level, real GDP and employment**. Some domestically produced goods that are protected may be used as inputs in the production of other goods. For example, if a textiles industry receives protection from textile imports, the domestic price of textiles will increase; clothes manufacturers that need textiles to produce clothes face higher costs of production. Using the *AD-AS* model we can see that short run aggregate supply decreases, the *SRAS* curve shifts to the left, causing cost-push inflation. What has happened is that real GDP falls and unemployment increases, while the price level also increases. The result is that final clothing products sell at a higher price.

- **Trade protection may have negative effects on a country's export competitiveness**. Some domestically produced goods that are protected may be used as inputs in the production of other goods that are exported. In the above example, if the products of the clothes manufacturers are exported, there will be lower competitiveness in export markets due to the higher price of clothes. Similarly, if a fertiliser industry is protected from fertiliser imports, the domestic price of fertiliser will be higher than the world price; farmers buying the fertiliser face higher costs of production, causing the final agricultural products to sell at a higher price. This results in lower competitiveness in export markets.
- **Trade protection may give rise to trade wars through retaliation**. As one country imposes barriers on imports, other countries may retaliate by imposing their own barriers. This can produce chain reactions with countries becoming more and more protectionist, with serious negative effects on global output and resource allocation.
- **Trade protection creates a potential for corruption**. For example, restrictions on imports may pave the way for bribes and smuggling goods illegally into a country, or may result in tariff and other revenues going into the pockets of bureaucrats rather than the government budget.

TEST YOUR UNDERSTANDING 15.1

1. a Identify some arguments in favour of trade protection.

   b Explain possible special circumstances under which they are considered to be valid.

   c What problems can they give rise to?

2. Using examples of barriers to trade, discuss some arguments against trade protection.

3. a Explain the infant industry argument.

   b What is its theoretical justification?

   c Explain some potential problems it may give rise to.

REAL WORLD FOCUS 15.2

The US wages a trade war

Beginning in January 2018, the United States began imposing a series of tariffs on various goods imported from several countries. Following the imposition of US tariffs, the affected countries began to retaliate.

A fundamental reason behind the tariffs was the US trade deficit (an excess of imports over exports; see Chapter 16) which is the largest in the world. According to this argument the imposition of tariffs would reduce imports, thus helping to correct the deficit. In addition, it was believed that tariffs would help protect jobs in the manufacturing sector.

Figure 15.2: The US imposes tariffs on imports

Part 1: Tariffs on steel and aluminum imports

In March 2018, the United States imposed tariffs of 25% on steel imports and 15% on aluminum imports. The rationale was that these tariffs were needed on the grounds of national security. In fact it was stated that the steel and aluminum imports 'threaten to impair the national security' of the United States. Yet most US imports of steel and aluminum are produced by US allies such as Canada and the European Union.

CONTINUED

Part 2: US trade war: effects on US employment

In his presidential campaign, President Trump promised to protect American jobs. Therefore the imposition of the tariffs was expected to lead to increases in US employment.

Yet by September 2018, the US Tax Foundation had estimated that the tariffs that had already come into effect plus planned tariffs together would lead to a loss of 459 816 jobs.

This could well be an underestimate. According to an analysis by Trade Partnership Worldwide, there would be an increase of 26 280 jobs in the steel and aluminum industries over the first one to three years, but employment throughout the rest of the economy would fall by 432 747, giving a net loss of 400 445 jobs. It was found that '16 jobs would be lost for every steel/aluminum job gained'.

Further, the US Center for Automotive Research claimed that if the United States imposes tariffs on cars, the loss of jobs would amount to 715 000. A study by the Federal Reserve (the United States central bank) in December 2019 notes 'We find that the 2018 tariffs are associated with relative reductions in manufacturing employment and relative increases in producer prices . . . For manufacturing employment, a small boost from the import protection effect of tariffs is more than offset by larger drags from the effects of rising input costs and retaliatory tariffs.'[8]

Reasons behind the job losses include higher costs of imported inputs, such as parts made of steel and aluminum, which make production less profitable. In some cases it is even possible for US firms to relocate to other countries in order to avoid the price increases due to the tariffs (see Part 3 below). Moreover, tariffs on US imports imposed by other countries in retaliation have the effect of reducing US exports, further hurting US firms.

Part 3: US trade war with China: effects on the car market

One of the countries most strongly affected by the US tariffs is China, which accounts for the largest share of the US trade deficit.

China responded to the US tariffs by quickly retaliating with a series of its own tariffs on billions of dollars worth of US goods. One category of these goods was cars. In the summer of 2018, it increased its tariff on US cars from 25% to 40%. But in an effort to boost domestic consumption it reduced its tariffs on car imports from all other countries to 15%. In addition, it lowered tariffs on 1400 products including household goods, apparel and appliances.

Overall imports from South Korea, Japan and the European Union increased by 31%, 24% and 20% respectively in the summer of 2018 while imports from the United States increased by a mere 11%. At the same time, due to strong consumer demand, Chinese exports to the United States increased by 11%, in spite of the US tariffs on Chinese goods.

American auto makers are very concerned about the impacts of the Chinese tariff. Volvo, which had recently opened a new plant in South Carolina with a promise of 4000 new jobs, has stated that it may have to go back on its promise in view of the US steel tariffs and threats of retaliatory tariffs from other countries.

BMW, the largest US car exporter, is already expanding its production in China, moving production away from South Carolina in the United States. This way it will be able to avoid the high tariff on US-made cars. Ironically, jobs in the United States are not increasing, but they might actually be moving to China.

Sources: *https://asia.nikkei.com; www.nytimes.com; www.forbes.com;*

[8] https://www.wbtv.com/2019/12/30/fed-study-trump-tariffs-caused-job-losses-higher-prices/

CONTINUED

Applying your skills

1 What do you think of the claim that steel and aluminum tariffs are justified on national security grounds when most imports come from US allies? What kind of trade protection do you think this is?

2 Use a tariff diagram to explain why tariffs were expected to give rise to new jobs in the United States.

3 Using an *AD-AS* diagram, explain why the US tariffs are likely to lead to job losses in the overall economy.

4 Using the same diagram as in question 3, and taking into account possible effects on the price level, explain the effects of the tariffs on US consumers. What kind of inflation is involved here?

5 Use a demand and supply diagram, to illustrate and explain the effect that retaliatory tariffs by other countries will likely have on the demand for US goods.

6 Using an *AD-AS* diagram, explain the likely effects on economic growth of countries engaged in trade wars.

7 Research current news reports on the impacts of the tariffs on

   a unemployment,

   b the price level, and

   c various product markets such as cars, outlining the most recent effects of the tariffs.

15.2 Economic integration: trading blocs

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between preferential trade agreements: bilateral, regional and multilateral – WTO (AO1)
- explain and provide examples of different types of trading blocs (AO2)
  - free trade area / agreement
  - customs union
  - common market
- discuss the advantages and disadvantages of economic integration arising from the different types of trading blocs (AO3)

Economic integration refers to economic co-operation between countries and co-ordination of their economic policies, leading to increased economic links between them. It occurs because the co-operating countries expect to derive benefits from policy co-ordination. It often begins by agreement between countries to reduce or eliminate trade and other barriers between them, and can extend to co-operation on other matters, such as labour policies or environmental policies and even monetary policy. There are various degrees of integration, depending on the type of agreement and the degree to which barriers between countries are removed.

Trade agreements and trading blocs

Preferential trade agreements

A **preferential trade agreement** (PTA) is an agreement between two or more countries to lower trade barriers on particular products in trade between each other. Trade barriers may remain on the rest of the products, and on imports from non-member countries. The result is that a member of the agreement has easier access to the markets of other members for the selected products, than countries that are not members.

PTAs sometimes involve co-operation between members on other issues, such as labour standards, environmental issues or intellectual property laws.

PTAs can take several forms, including free trade areas, customs unions or common markets, and they may be bilateral (involving two countries) or regional (involving several countries).[2]

Bilateral, regional and multilateral (WTO) trade agreements

A **bilateral trade agreement** is an agreement between two countries, whereas a **multilateral trade agreement** involves an agreement between many countries. Another distinction involves **regional trade agreements**, which as the term suggests involves trade agreements between a group of countries that are within a geographical region. The main objective of bilateral, regional and multilateral trade agreements is to promote **trade liberalisation**, which is free (or freer) trade by reducing or eliminating trade barriers between members.

The trade agreements reached under the World Trade Organization (WTO) are *multilateral*, because they include many member countries around the world (164 in 2019) and because they require all member countries to reduce trade barriers at the same time. One of the fundamental principles of the WTO is *non-discrimination*, meaning that a country cannot discriminate between any WTO members. In other words, it cannot impose higher barriers on imports from one country and lower ones on imports from another country. *This is a fundamental principle for the development of free trade globally*. However, the WTO makes an exception for bilateral and regional trade agreements, even though all preferential trade agreements involve discrimination against non-member countries.

Trading blocs

A **trading bloc** is a group of countries that have agreed to reduce tariff and other barriers to trade for the purpose of encouraging free or freer trade and co-operation between them. Beginning with the lowest degree of economic integration, we can distinguish between the following trading blocs.

Free trade area (agreement)

A **free trade area (agreement)** consists of a group of countries that agree to gradually eliminate trade barriers between themselves, and is the most common type of integration area. Each member country retains the right to pursue its own trade policy towards other non-member countries (to impose its own trade barriers). In trade relations between members, there may be free trade in some products, and some protection in other products.

Examples of free trade areas (FTAs) are NAFTA (North American Free Trade Agreement), including Canada, Mexico and the United States; and ASEAN (Association of Southeast Asian Nations), and SAARC (South Asian Association for Regional Cooperation).

One problem that arises in free trade areas is that a product may be imported into the FTA by the country that has the lowest external trade barriers, and then sold to countries within the FTA that have higher external trade barriers. This problem arises because each country has its own individual barriers toward non-members. It creates difficulties for those countries with higher barriers because they may end up importing more of the good than they would like. To deal with this problem, free trade areas make complicated 'rules of origin' for imports, designed to prevent goods from entering countries with lower external barriers.

Customs union

A **customs union** consists of a group of countries that fulfils the requirements of a free trade area (elimination of trade barriers between members) and in addition adopts a common policy towards all non-member

[2] 'Preferential trade agreement' is sometimes used in another sense to refer to the weakest form of economic integration, coming **before** the formation of bilateral or regional trading blocs. This type of PTA is not of interest to us because it is not allowed by WTO rules.

countries. Each country in a customs union is no longer free to determine its own trade policy towards non-member countries. Also, the member countries of the customs union act as a group in all trade negotiations and agreements with non-members. A customs union therefore involves a higher degree of economic integration than a free trade area.

Examples of customs unions include CEFTA (Central European Free Trade Agreement), SACU (South African Customs Union), PARTA (Pacific Regional Trade Agreement), and others.

Customs unions have the advantage over FTAs that they avoid having to create complicated 'rules of origin' for imports, since they all have the same common external barriers. However, customs unions face the problem that they must co-ordinate their policies toward non-members. This gives rise to the possibility of disagreements, as they may not all agree on what are appropriate levels of tariff and other barriers on non-members.

Common market

A **common market** is an even higher degree of economic integration, in which countries that have formed a customs union proceed further to eliminate any remaining tariffs in trade between them; they continue to have a common external policy (as in a customs union), and in addition, they agree to eliminate all restrictions on movements of any factors of production within them. The factors of production of importance are labour and capital, which in a common market are free to cross all borders and move, travel and find employment freely within all member countries.

The best-known common market is the European Economic Community (EEC, the precursor of the present European Union), formed in 1957. Another example is the Caribbean Community (CARICOM) Single Market and Economy (CSME).

A common market offers major advantages to its members compared to FTAs and customs unions. They enjoy free trade and all its advantages (lower prices, greater consumer choice, etc., see Chapter 14). Workers are free to move and work in any member country without restrictions, and capital (physical and financial) can also flow from country to country without restrictions. This results in a better use of factors of production. For example, there may be high unemployment in one country, and a high demand for labour in another country. This will encourage unemployed workers to seek work in the country facing labour shortages. Similarly, if the profitability of investing is greater in one country than in another, capital will gravitate to the more profitable country, making better use of capital resources. Factor mobility across countries improves the allocation of resources.

However, the development of a common market requires even greater policy co-ordination among members than in a customs union, and requires the willingness of member governments to give up some of their policy-making authority to an organisation with powers over all the member governments. Both these requirements can be difficult to accomplish, and need a long time for all countries to make the necessary policy changes to achieve co-ordination. For this reason there are far fewer common markets in the world than free trade areas and customs unions.

Evaluating trading blocs

Possible advantages of trading blocs

Economic integration over the long term can be expected to bring forth many of the benefits of free trade. Benefits that member countries can expect to derive include the following:

Trade creation **(HL only)**

When a trading bloc is established, patterns of trade between countries change, since trade between the members is encouraged through the lowering of trade barriers, while trade with non-members is discouraged through the maintenance of trade barriers. **Trade creation** refers to the situation where higher cost products (imported or domestically produced) are replaced by lower cost imports.

Consider an example. Suppose Cottonia and Microchippia both produce cotton. Cottonia has a comparative advantage in cotton, it has a lower cotton price, and therefore Microchippia imports cotton from Cottonia. Initially, Microchippia imposes tariffs on its cotton imports, this way protecting its own cotton producers. Then Cottonia and Microchippia form a bilateral trade agreement, and tariffs on cotton are abolished. Microchippia's cotton imports increase (corresponding to an increase in Cottonia's cotton exports), and its domestic production of cotton decreases. This is a case of *trade creation*, because higher cost domestic cotton production in Microchippia is partly replaced by lower cost imports of cotton.

The benefits of trade creation include getting rid of disadvantages of tariffs. The decrease in Microchippia's domestic production of cotton leads to greater efficiency in production, and together with the increase in consumption made possible by more imports, there is greater allocative efficiency.

In general, trade creation has the effect of increasing social welfare.

- ***Increased competition***

The removal of trade barriers within trading blocs results in increased competition among producers in member countries. With low or no barriers, imports increase, forcing domestic producers to compete with lower cost producers from other countries. Trade barriers, on the other hand, protect inefficient domestic producers. Increased competition offers major advantages in terms of production by more efficient producers, lower prices for consumers and improved allocation of resources.

- ***Expansion into larger markets***

This is an obvious benefit arising from the ability of firms to sell beyond their national boundaries, and increasing their exports.

- ***Economies of scale***

This follows from the above point. In a small market a firm cannot take advantage of economies of scale (lower average costs) since a firm cannot grow large enough so that its average costs begin to fall substantially. When an economy opens itself up to free trade with other countries, its exports are likely to increase and as the size of the market expands, the firm can achieve lower average costs.

- ***Lower prices for consumers and greater consumer choice***

The elimination of trade barriers (along with increased competition and economies of scale) results in lower prices for consumers. In addition, increased imports mean a greater of variety of goods from which consumers can choose.

- ***Increased investment***

Enlarged markets often give rise to increased investment by firms that want to take advantage of the larger market size. This investment may be internal, that is, by firms originating from a country within the trading bloc, or external, originating from a country that is outside the trading bloc (by multinational corporations). A major incentive faced by outsider firms to set up production units within the bloc is that they escape the tariff or other protection that the trading bloc imposes on imports from outside. One of the incentives faced by countries to form trading blocs is to attract investments by multinational corporations (to be studied in Chapter 20).

- ***Better use of factors of production: improved resource allocation and greater employment opportunities***

If a trading bloc develops into a common market, which involves free movement of factors of production, specifically capital and labour, there will also be a better use of these within the bloc. As discussed above, unemployed workers in one country may seek a job elsewhere where there are more employment opportunities. Capital can also move freely in search of greater profits. A better allocation of resources results.

- ***Improved efficiency in production and greater economic growth***

As we know from Chapter 14, the elimination of trade barriers and free trade lead to improved efficiency in production, as inefficient producers lose their protection, leading to better prospects for achieving more rapid economic growth.

- ***Stronger bargaining power***

When countries bargain individually in multilateral negotiations, such as with the World Trade Organization, they do not have much bargaining power especially if they are relatively small. If they bargain as a trading bloc they have much greater power, increasing their chances of being heard and achieving their objectives.

- ***Political advantages***

Greater economic integration is likely to result in a reduced likelihood of hostilities arising between countries whose economies are becoming more interdependent through increased trade, investment, labour and financial flows. Further, economic integration may lead to political stability as well as co-operation, resulting in further benefits for member countries.

Possible disadvantages of trading blocs

Trade diversion (HL only)

Trade diversion refers to the situation where lower cost imports are replaced by higher cost imports from a member after the formation of the bloc. It is the opposite of trade creation,

Suppose Cottonia, Robotia and Microchippia all produce cotton. Cottonia is the lowest cost producer of the three, followed by Robotia, and then by Microchippia, which is the highest cost producer. Initially, Microchippia imposes a tariff on all imports of cotton, regardless of country of origin. Since Cottonia is the lowest cost producer, Microchippia imports from Cottonia and not from Robotia (Cottonia's cotton price plus the tariff is lower than Robotia's price plus the tariff). Microchippia then decides to form a trading bloc with Robotia. It therefore eliminates the tariff on cotton from Robotia, and maintains the tariff on cotton from Cottonia. The result is that it now becomes cheaper for Microchippia to import cotton from Robotia rather than Cottonia. Microchippia's imports have shifted from a lower cost producer, Cottonia, to a higher cost producer, Robotia; this is therefore a case of trade diversion.[3]

The possibility of trade diversion resulting from a trading bloc is an argument *against* trading blocs, and *in favour* of multilateral (WTO) trade liberalisation. The reason is that trade diversion cannot occur with multilateral reduction or elimination of trade barriers. Trade diversion occurs when an importing country is forced to import from a higher cost producer within a trading bloc, whereas before it joined the trading bloc it was importing from a lower-cost producer elsewhere. If all countries reduce their barriers at the same time, it is not possible for lower-cost imports to be replaced by higher-cost imports; the importing country will simply import from lower cost producers who sell at lower prices.

While trade creation, discussed above, has the effect of increasing social welfare, trade diversion reduces it. Therefore, whereas a trading bloc creates free trade for the members, it may or may not improve the allocation of resources. Resource allocation will improve only if trade creation effects are larger than trade diversion effects.

Note, however, that even if a trading bloc leads to trade diversion over the short term, it is possible that the longer-term positive effects discussed above will more than compensate countries for possible short-term losses. According to some studies, the long-term effects may be five or six times more important than the short-term ones.

- ***Trading blocs may be a challenge to multilateral (WTO) trading negotiations (trade liberalisation)***

Many economists believe that while the establishment of trading blocs with free trade between members may be an improvement over trade protection, trading blocs are inferior to the WTO's multilateral approach of reducing trade barriers towards all countries. They believe that the break-up of the world trading system into many blocs can create trade conflicts between different blocs that may slow down the process of global trade liberalisation. Some large trading blocs may enjoy free trade and all its benefits within the bloc, but may impose high trade barriers on non-members, and the result may be to limit trade rather than to increase it on a global scale. This could lead to a worse global allocation of resources.

- ***Unequal distribution of gains and possible losses***

Countries forming a trading bloc are unlikely to gain equally from the operation of the trading bloc, and this creates the potential for conflicts between the members and makes it difficult to reach agreements. It is also possible for some countries to gain while others become worse off in some respects. The same applies to gains and losses within the member countries, as some stakeholders are likely to gain while others lose. This is an important issue that we will come back to when we evaluate preferential trade agreements as a trade strategy to achieve economic growth and development in Chapter 20.

- ***Economic integration involves a loss of sovereignty***

Sovereignty refers to authority over decision-making within the national economy. The formation of trading blocs involves giving up some domestic decision-making power to a supranational authority. The loss of sovereignty is the least in a free trade agreement, becoming progressively greater in the case of a customs union and common market. The loss of sovereignty is even greater in the case of monetary union (see below).

TEST YOUR UNDERSTANDING 15.2

1 a Define and distinguish between a free trade area, customs union and common market.

  b How do these illustrate an increasing degree of economic integration?

2 Research and find real-world examples of free trade areas, customs unions and common markets. Choose one or more of these and discuss some of the advantages and disadvantages.

3 What do you think is one reason why countries that want to form a trading bloc usually start out by forming a free trade area, and then gradually move towards a customs union, and even later towards a common market?

[3] A full analysis of trade diversion is actually a little more complicated, because trade diversion comes with some benefits, however this discussion is beyond the scope of this book.

REAL WORLD FOCUS 15.3

African nations form AfCFTA

Figure 15.3: Kigali, Rwanda. The African heads of state establishing the African Continental Free Trade Area (AfCFTA)

In March 2018, after five years of planning, the African Continental Free Trade Area (AfCFTA) was launched at the Kigali Summit (in Rwanda). With 55 countries, a population of 1.27 billion and GDP of $2.5 trillion, it will be the largest free trade area in the world.

Its objectives are to create a single market for goods and services in the entire African continent, together with free movement of labour and capital, thus eventually becoming a common market. It is expected to lead to major cost advantages for firms that will have opportunities to benefit from economies of scale and increased competitiveness. According to forecasts, trade within Africa will increase by 52% by 2022 relative to 2010. This is significant because trade within Africa (intra-Africa exports) in 2017 were only 16.6% of total exports, compared with 68% in Europe and 59% in Asia.

In addition, AfCFTA aims to achieve sustainable and inclusive development, to promote gender equality, and to promote industrial development through diversification. According to the Director of the African Export Import Bank (Afreximbank), 'Constrained access to markets limits the growth of firms. Therefore, for domestic firms, getting rid of local market constraints may improve growth prospects and access to finance and technology in the global economy. There are, however, notable challenges. If large firms gain a dominant position in the African market, they may crowd out small and medium-sized firms.'

Small- and medium-sized firms, which absorb more than 80% of Africa's employment, are expected to benefit from new markets, reductions in input costs, increased efficiency and increased sales.

However, there are also challenges. Some small- and medium-sized firms may be unable to withstand the competition from larger businesses. Trade liberalisation may hurt the poor. Workers from poor countries may work long hours and live in conditions of poverty in order to send money home to their families. Stronger competition may lead to greater environmental degradation as small- and medium-sized firms try to cut costs by dumping wastes.

Moreover, African countries vary enormously by level of economic development and by size. Three countries alone, Egypt, Nigeria and South Africa contribute over 50% of Africa's GDP while six island nations together contribute just 1%. The weaker countries may require preferential treatment in order to ensure that the risks arising from increased competition are reduced for them. For example, they could be permitted to maintain tariffs or other forms of protection for infant industries or for the purposes of diversifying their economies.

According to the Nigerian Labour Congress (an organisation of trade unions), the trade agreement is an 'extremely dangerous and radioactive neo-liberal policy initiative'.

On the other hand, it is also argued that with all of Africa united as a huge bargaining unit, it can be far more influential in international trade negotiations with other countries or groups of countries around the world. According to Professor Ngaire Woods at Oxford University, "Africa could stride onto the trade negotiation stage as one enormous market. This could lead to a new engine of growth across the continent."[1]

Sources: *www.weforum.org/agenda; www.sunnewsonline.com; https://www.weforum.org/agenda/2019/09/africa-just-launched-the-world-s-largest-free-trade-area*

CONTINUED

Applying your skills

1 Describe how AfCFTA represents economic integration. Identify what feature will allow it to evolve into a common market.

2 Explain what AfCFTA member countries hope to gain from integration.

3 Identify the stakeholders who are likely to gain as well as those likely to lose from the formation of AfCFTA.

4 Why are some of the member countries' fearful of integration? Identify and explain their concerns.

5 (HL only) Suppose one AfCFTA member country has an absolute advantage over another AfCFTA member country in the production of all goods.

  a Using a diagram, explain under what conditions it may still be possible for both countries to gain from trade.

  b Under what circumstances might specialisation according to comparative advantage not be desirable? Discuss.

6 Using the information in the text, evaluate the likely effects of AfCFTA on the economies of the member countries.

REAL WORLD FOCUS 15.4

Brexit: The UK leaves the European Union

Figure 15.4: London, United Kingdom. Pro-Brexit supporters gather in Parliament Square

Historical background: the development of the European Union

The European Union has its origins in the European Coal and Steel Community (ECSC), a free trade area formed in 1952 by a small group of six European countries.[4] The ECSC abolished trade barriers in important military and economic resources including coal, steel, scrap and iron ore. At the same time, it established institutions for administration, legislation and policy as well as a Court of Justice to resolve disputes. The creation of this free trade area was the first step in a series of efforts toward closer integration of European countries. One of the key motivating factors was to rebuild Europe after the Second World War as well as to prevent future hostilities and promote lasting peace between France and Germany.

[4] The six countries were Belgium, France, Italy, Luxembourg, Netherlands and West Germany.

CONTINUED

In 1957, the same group of countries formed the European Economic Community (EEC), which included lowering of trade barriers for a far broader range of products. The United Kingdom joined the EEC in 1973. The EEC established a common market which in addition to free trade and common trade policy toward non-members also included free movement of labour and capital, a Common Agricultural Policy (CAP), a social policy to promote labour mobility and worker protection, as well as policies to promote market competition, including regulations preventing anti-competitive behaviour by firms. Its membership grew gradually over the years, and by 2013 there were 28 members.

The above policies, involving increasing economic integration between the member countries, were implemented by supranational bodies with authority over national governments. There are several such bodies, the most important of which are the European Commission, which proposes new policies and legislation; the Council of the European Union and the European Parliament, responsible for legislation (laws) that all member countries must abide by; and the European Court of Justice, that resolves disputes.

In 1993, the EEC changed its name to the European Union, known as the EU. Over time, additional functions and objectives were developed, among which were co-ordination of foreign policy, development assistance policy for less developed countries, a common research and development policy (R&D), a social policy protecting workers rights and a regional development policy that supports training to reduce unemployment.

Eventually 19 of the 28 EU countries gave up their national currencies, adopted the euro as their currency, and gave up their own independent monetary policy through the establishment of the European Monetary Union (see below). *The United Kingdom was not among the countries that adopted the euro*, preferring to be in the EU with its own currency, the British pound, and its own independent monetary policy.

Gradually, over the decades, the European Union has achieved an unprecedented degree of economic integration, far greater than any other international organisation.

Brexit: UK decides to leave the European Union

In June 2016, after 43 years of membership, the United Kingdom held a referendum in which its citizens were asked if they wished the United Kingdom to leave or remain within the European Union. England and Wales voted to leave, while Northern Ireland and Scotland voted to remain. The result overall came out in favour of leaving with 52% 'Leavers' as opposed to 48% 'Remainers'. This event came to be known as *Brexit*, made up of the words Britain and exit.[5] It was received with great surprise around the world as the expectation everywhere was that the United Kingdom would choose to remain in the EU.

The main arguments in favour of leaving the EU include the following:

- ***Loss of sovereignty*** Many Leave voters believe that decisions that involve the United Kingdom should be taken in the UK by the UK government, not by bureaucrats in Brussels where the EU's administration is located. It is argued that the EU institutions are not accountable to voters.
- ***Opposition to immigration*** The *free movement of labour* principle of the EU results in net inward migration into the United Kingdom from other EU countries. Many Leavers believe that the United Kingdom should take back control of its borders so as to regain the ability to restrict immigration, and control the range of skills of workers who enter the country in accordance with needs.

[5] Britain is composed of England, Scotland, Wales. The United Kingdom on the other hand is composed of the three countries of Britain plus Northern Ireland. The full name is United Kingdom of Great Britain and Northern Ireland.

CONTINUED

- ***Opposition to the customs union*** As a member of the EU customs union, the United Kingdom is unable to negotiate trade agreements with third countries independently, as it must conform to the common external policy toward third countries of the EU. Leaving the EU would mean that the United Kingdom can negotiate any trade deal it wishes with any other country or group of countries.
- ***Opposition to contributions to the EU budget*** The EU collects money from each of the member countries, which is spent on the Common Agricultural Policy, support for depressed regions, and other activities. EU membership has a net cost to the UK of 7.1 billion pounds annually, which amounts to about 0.03% of GDP according to the Office of National Statistics. Opponents to the EU, or Leavers, argue that the UK should take back control over how this money is spent.
- ***Opposition to the euro*** The United Kingdom has not adopted the euro as noted above. However, numerous economists argue that the survival of the euro depends on the euro zone countries becoming even more economically integrated such as through a fiscal union. As Leavers are opposed to more integration and loss of sovereignty, they see the euro as a threat.
- ***Opposition to the Common Agricultural Policy (CAP)*** The CAP, which supports EU farmers though a range of interventionist measures such as price controls and subsidies, absorbs a large proportion of EU funds. Yet it promotes inefficiency. Leavers argue that the United Kingdom does not need support of the EU's CAP nor do they want to contribute to its funding.
- ***Opposition to legislation protecting workers*** It has been argued that EU legislation that protects workers' rights is costly, that very significant savings could be made by getting rid of social and employment protection.
- ***An argument from the political right: the EU imposes too many regulations*** Some people argue that the EU has too many rules and regulations as well as social protection (see above) that stifle economic activity and private initiative, while promoting left wing principles.
- ***An argument from the political left: the EU supports big corporations*** People on the opposite end of the political spectrum argue that the EU gives too much power to large corporations, supporting the interests of powerful elites.

The main arguments in favour of remaining in the EU include the following:

- ***Trade with the EU*** Roughly 50% of the UK's exports are toward the EU, with markets of over half a billion people, to which the United Kingdom has free access with no trade restrictions. If the United Kingdom leaves the EU, it will be very difficult if at all possible to negotiate trade deals with other countries that could replace the trade lost with the EU. The result would be a significant decline in UK exports.
- ***Investment*** Much of UK investment is linked with the guaranteed markets of the EU that the United Kingdom exports to. Leaving the EU and the potential loss of the EU markets poses a risk that UK investment will suffer.
- ***Foreign direct investment (FDI)*** The United Kingdom attracts a lot of foreign direct investment because it offers free access to the markets of the EU countries. Leaving the EU creates the risk that FDI will leave the United Kingdom, going to EU countries that maintain that access.
- ***Jobs and employment*** The above risks that arise from trade and investment create risks for the labour market that would suffer from a reduction in available jobs.
- ***Carrying out business in EU countries*** Common rules for carrying out business mean that there is no red tape and national regulations that UK firms must confront in their dealings with other businesses in the EU countries.
- ***EU worker legislation*** The EU has laws about equal pay for men and women; four weeks paid annual leave for all workers; bans on discrimination due to age, race or sexual orientation; plus numerous other benefits for workers.

CONTINUED

- ***The issue of immigration*** Some supporters of EU membership argue that immigration is a rich source of culture that should be welcomed. It is moreover an important source of labour skills of which there are not enough in the United Kingdom. Without immigrants there would be shortages of labour in the hospital and care services, as well as building and service industries. Moreover, nearly 750000 UK citizens live or work in other EU countries. UK citizens would lose the right to move to an EU country to live and work there.
- ***The Common Agricultural Policy (CAP)*** The CAP is in the process of being reformed, with a view to reducing its inefficiencies. If the United Kingdom remains in the EU it could participate in its reform which could work to the United Kingdom's long-term interest.
- ***The fall of the pound*** A Leave vote in the referendum was expected to result in a sharp drop in the pound due to the uncertainties that Brexit would create in the business world. In fact the pound did suffer a precipitous decline, which created inflationary pressures. Since the referendum, the UK has experienced inflation, and with nominal incomes stagnant, real incomes have been falling.
- ***Global political influence*** It is likely that the United Kingdom can have a greater influence in global politics and negotiations as a member of a large a powerful bloc like the EU, rather than acting on its own.

The referendum of June 2016 did not mean that the United Kingdom would immediately leave the EU. In March 2017, the British Prime Minister formally informed the EU that the United Kingdom was leaving (through the so-called Article 50). Following this, the United Kingdom had a period of two years to negotiate with the EU the nature of its relationship with the EU after its formal departure, scheduled to take place in March 2019. Due to inability to complete the negotiations, this date was later extended to 31 October 2019. Following another delay Brexit took place on 31 January 2020.

It became apparent during the negotiations that the United Kingdom wanted to maintain free trade with the EU, from which it has tremendous benefits, but with restrictions on immigration. The EU on the other hand refused to split the free movement of goods, services, people and capital.

Applying your skills

1. Distinguish between membership in the European Union and membership in the European Monetary Union.
2. Using an appropriate diagram explain why the drop in the pound created inflationary pressures in the United Kingdom.
3. Explain the difference between nominal and real incomes. Using the information in the text explain why real incomes in the United Kingdom have been falling.
4. Up to the time of Brexit, numerous studies had tried to predict the economic effects that this would have. Most of these were pessimistic but highly variable, predicting mildly negative to very strongly negative effects. Investigate the effects that Brexit has had on the UK economy by examining
   a. the value of the pound in relation to the euro,
   b. exports,
   c. the current account,
   d. unemployment,
   e. inflation,
   f. economic growth, and
   g. foreign direct investment.

15.3 Economic integration: monetary union

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain monetary union (AO2)
- discuss advantages and disadvantages of monetary union (HL only) (AO3) (This will be discussed in Chapter 17, Section 17.3 after you have studied exchange rates.)

Monetary union

Tip: Revisit this section after you have studied exchange rates in Chapters 16 and 17 as you will then be able to better understand the material in this section. You will be reminded in Chapter 17 to return to this section.

Monetary union involves a far greater degree of integration than a common market, and occurs when the member countries of a common market adopt a common currency and a common central bank responsible for monetary policy. A monetary union has been formed by a number of the countries of the European Union, known as the 'euro zone countries'. There are many other trading blocs around the world that have plans to form a monetary union in the future.

Following years of preparation for monetary union, 11 countries of the European Union adopted a single currency, the euro, in 1999. These countries were Austria, Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain. They were joined by Greece in 2001, Slovenia in 2007, and Cyprus and Malta in 2008. The countries that have adopted the euro are members of the European Monetary Union. On 1 January 1999 the new currency, the euro, came into being and the currencies of the participating countries were locked together through fixed and absolutely unchangeable exchange rates. Euro notes and coins were introduced on 1 January 2002 and for the period of one year they coexisted side by side with national currencies. On 1 January 2003 the national currencies of the participating countries were abandoned and the euro became the sole currency of the euro zone countries.

The creation of the European Monetary Union was one of the most significant economic events in the post-Second World War period, as it was the first time ever that such a large group of countries gave up their national currencies to adopt a single common currency. As part of their preparation for membership, the countries had to agree to a number of conditions known as 'convergence requirements', including limiting their rate of inflation, limiting their budget deficit to 3% of GDP, limiting their government debt to 60% of GDP, and others. Moreover, in adopting the common currency, they gave up a significant part of their economic sovereignty to a supranational body, the European Central Bank, which assumed the responsibility for monetary policy for all the member countries. Following the adoption of the euro, the member countries gave up control of their money supply and their ability to carry out their own monetary policy, transferring these powers from each of their national central banks to a single institution, the European Central Bank.

Monetary union, or the formation of a single currency, can be partly thought of as a system of 'fixed' exchange rates among the participating currencies, but one in which there is no possibility ever of changing the value of one currency in relation to another (no possibility of ever revaluing or devaluing, see Chapter 16).

TEST YOUR UNDERSTANDING 15.3

1. Explain the meaning of monetary union.
2. Outline why it is a higher form of integration than a common market.

15.4 World Trade Organization

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the objectives, functions and factors affecting the influence of the World Trade Organization (WTO) including difficulties in reaching agreement and members' unequal bargaining powers (AO2)

WTO functions and objectives

The **World Trade Organization (WTO)** describes itself as an organisation for liberalising (freeing up) trade, with the following objectives and functions:[6]

- ***It administers WTO trade agreements***. The WTO helps in the implementation and administration of international trade agreements.
- ***It provides a forum for trade negotiations***. The WTO provides a forum for members to discuss their trade problems and negotiate trade agreements on how to liberalise trade. This is one of the most important WTO functions.
- ***It handles trade disputes***. When WTO members disagree on trade issues, the WTO makes decisions to resolve the differences on the basis of the legal foundations of the trade agreements.
- ***It monitors national trade policies***. The WTO carries out periodic reviews of its members' national trade policies. Members are required to notify the WTO of any changes in trade policy. The WTO also examines new trading bloc arrangements.
- ***It provides technical assistance and training for developing countries***. The WTO provides assistance in the form of training of government officials as well as journalists, academia and private sector representatives in developing countries on trade-related issues arising from WTO trade agreements.
- ***It facilitates co-operation with other international organisations***. The WTO co-operates with other international organisations (such as the World Bank and International Monetary Fund discussed in Chapter 20), in order to facilitate co-ordination of global polices.

The WTO's influence in international trade: criticisms and challenges

The WTO claims to offer benefits to the global trading system arising from its contribution to the development of free trade, the establishment of an effective system of trading rules and a mechanism to resolve trade disputes between countries. Yet the WTO is a highly controversial organisation in the world today, rousing passionate feelings among both supporters and critics.

The WTO is accused of promoting trade rules that do not favour developing countries

The most comprehensive and far-reaching trade liberalisation agreement in history, the Uruguay Round, was achieved in 1994 under the GATT, the WTO's predecessor. It resulted in tariff reductions of 33% on average, with the share of goods with no tariffs increasing from 20–22% to 40–45%. Yet the agreements did not succeed in providing countries with equal opportunities to share in the benefits of increased trade The 1997 Human Development Report of the United Nations Development Programme (UNDP) summarises the impacts of the Uruguay Round.

> *'Poor countries often lose out because the rules of the game are biased against them, particularly those relating to international trade. The Uruguay Round made little difference. Although developing countries have three-quarters of the world's people, they will get only a quarter to a third of the income gains generated and most of that will go to a few powerful exporters in Asia and Latin America.'*[7]

[6] http://www.wto.org/english/thewto_e/whatis_e/tif_e/fact1_e.htm

[7] United Nations Development Programme, *Human Development Report* 1997 (available at http://hdr.undp.org/en/), p. 85.

REAL WORLD FOCUS 15.5

History and background to the World Trade Organization

Figure 15.5: Quezon City, Philippines. Protestors opposed to WTO membership believed to have led to 'massive flooding of imported food and agricultural products'

During the Great Depression of the 1930s, when countries around the world were suffering major declines in output and high rates of unemployment, they resorted to tariffs and other restrictions to limit imports and protect domestic production and employment. As each country raised its tariffs, it provoked retaliatory tariff increases from its trading partners, resulting in 'tariff wars' that greatly reduced the volume of international trade without positive effects on output and employment.

In 1947, 23 countries formed an agreement known as the General Agreement on Tariffs and Trade (GATT), intended to gradually liberalise (free up) international trade and prevent further outbreaks of tariff wars. Up to that time, all countries had been free to impose any type of trade protection measures on imports from other countries.

The GATT was based on the following principles:

- *non-discrimination*: equal, non-discriminatory treatment for all member countries; this means that a country's trade policy cannot discriminate between its trading partners (for example, it cannot impose a higher tariff on imports from one country and a lower one on imports from another country); exception was made for bilateral and regional trading blocs
- *elimination of non-tariff trade barriers:* gradual elimination of non-tariff barriers (quotas, subsidies, administrative barriers); exceptions were made for agricultural products and countries with balance of payments difficulties
- *consultations to resolve trade disputes:* GATT provided a forum for discussions to resolve disagreements between countries.

The members of GATT would periodically conduct 'rounds' of negotiations in which they tried to achieve the above objectives. They conducted eight rounds in total, the last one being the 'Uruguay Round' (because it was launched in Punte del Este in Uruguay), which lasted from 1986 to 1994. In 1994, by which time the member countries had increased to 124, an agreement was reached to replace the GATT by a new international trade body. This body was established on 1 January 1995, and was called the World Trade Organization (WTO).

Today, the WTO provides the institutional and legal framework for the trading system that exists between member nations worldwide. As of 2019, it had 164 members, accounting for over 98% of the value of global trade. In addition, it had 23 'Observer' countries, which must apply for full membership within five years of having Observer status.

Applying your skills

1 Explain the main purpose behind the creation of the GATT, which was the precursor of the WTO.

2 In view of increasing protectionist trends in the global economy, and the dangers posed by trade wars, discuss the potentially important role of the World Trade Organization in the global economy.

Reasons for the unfavourable treatment of developing countries included the following:

- Developed countries received greater tariff reductions than developing ones.
- The practice of making increased use of non-tariff and hidden barriers against developing country exports was not sufficiently addressed.
- Whereas the Uruguay Round provided for reductions in agricultural subsidies, these were not implemented because of resistance in developed countries.
- Protection of intellectual property increased the costs of acquiring new technology by developing countries.
- Multinational corporations (MNCs, see Chapter 20) no longer had to buy their supplies locally; this meant they could no longer increase demand for locally produced goods and services, which would increase local employment.

The WTO has been unable to reach an agreement on agricultural protection and services

In 2001, the WTO began negotiations known as the 'Doha Development Round' (launched in Doha, Qatar). However, it soon became apparent that developed country interests were dominating the negotiations. By the end of July 2008 the negotiations collapsed, mainly because of *the inability to reach agreement on protection of developed country farmers*. Developed countries have long been protecting their farmers through production and export subsidies, with numerous negative effects on the farmers and economies of developing countries (See Real world focus 14.2, Chapter 14 and Chapter 19).

While farmer protection continues in many developed countries, particularly the United States, the European Union and Japan, one key success of the WTO should be noted here, which is an agreement in 2015 to phase out export subsidies (Real world focus 14.2, Chapter 14).

The WTO is accused of not distinguishing between developed and developing economies

The WTO treats all countries as if they are at the same level of development, with the sole exception of *Least Developed Countries* (low income countries identified by the United Nations which face severe constraints in achieving sustainable development (to be discussed in Chapter 18)). This means that agreements requiring countries to lower their trade protection for the most part apply to developed and developing countries alike. However, there are many developing countries that may need trade protection on the grounds of developing their infant industries or diversifying their production and economies to reduce reliance on **primary products** (primary commodities).

The WTO is accused of ignoring environmental and labour issues

It is argued that the WTO does not pay enough attention to issues relating to the environment. For example, it has encouraged removal of trade barriers on imports from countries that have low environmental protection standards. It has also not tried to reduce subsidies on activities that harm the environment such as agriculture, coal and transport. Supporters of the WTO argue that this is the fault of the member countries which are not sufficiently interested in addressing environmental issues. Developed countries oppose the elimination of subsidies, while developing countries claim that WTO rules relating to the environment may be a form of 'hidden' protection.

Regarding labour standards, the WTO is accused of ignoring issues like child labour or other violations of internationally accepted labour standards. The WTO has not linked the removal of trade barriers to improved labour practices. Developing countries oppose the inclusion of labour standards in trade agreements, in part because they are afraid that this can also be used as a type of 'hidden' protection.

WTO members have unequal bargaining power

From a formal point of view, decision-making in the WTO is democratic because of the WTO rule that each member country has one vote.[8] Yet decisions are usually made by consensus, meaning that once there is a proposed decision, a consensus emerges if no member present at the meeting objects to the decision.

Critics argue that the WTO is actually undemocratic, and decisions are based on the power of members in spite of the one vote per member rule. Economically

[8] This is in contrast to the voting powers at the International Monetary Fund (IMF) and World Bank (see Chapter 20) where voting power is in accordance with the size of contributions to the budget, giving obvious powers to the larger and wealthier countries.

powerful countries dominate agenda-setting and express opinions that carry greater weight. Less powerful countries often remain silent and give in to the demands of the more powerful in fear of possible retaliation.

In addition, the difficulty of reaching a consensus in view of the large number of countries involved leads to the formation of country coalitions which increases the power of the wealthier countries. Another issue involves the process of negotiations, which includes meetings that often restrict participation to the more influential and dominant countries. This, combined with the sometimes weak negotiating abilities of weaker members, results in outcomes that favour the more powerful.

The silence of developing countries was an important factor in the outcomes of the Uruguay Round of 1994 (see above), which benefitted the developed countries. In view of the above issues, there have been many calls for reforming the WTO's decision-making process with a view to making it genuinely democratic and hence more effective.

A key challenge faced by the WTO: fragmentation of global trade

In addition to agreeing on the phasing out of export subsidies (see above), the WTO also succeeded in making an agreement between some of the WTO members involving cutting tariffs on information technology. Also, the WTO is credited with having prevented a return to full scale trade protection during the global financial crisis beginning in 2008.

In spite of these successes, there are concerns that the WTO has come to a stalemate created by developed country demands that developing countries open up their markets to industrial products and services, while they themselves continue to offer protection to their farmers. Moreover, there are fears that the global trading system may be facing a setback because of growing trade protection tendencies around the world, strongly reinforced by the position of the United States since 2018 (see Real world focus 15.2). In addition, growing impatience with the WTO as the main mechanism for trade liberalisation is a factor that has resulted in a significant increase in free trade agreements around the world. The number of trade agreements reported to the WTO grew from 20 in 1990 to 159 in 2007 and to 270 in 2017.

Another response on the part of countries impatient with the WTO has come in the form of *plurilateral agreements*, suggested by the US and Canada, which involve agreements by WTO members on a *voluntary basis*. Pluralistic agreements have been strongly criticised on the grounds that they will seriously undermine the multilateral approach of the WTO which involves participation in agreements by *all member countries*.

Another key challenge faced by the WTO: the blocking of its powers to resolve disputes

One of the most important roles of the WTO is its ability to resolve trade disputes between countries that disagree on trade practices. The WTO has been carrying out this function through its "Appellate Body", a committee of seven judges who hear complaints of WTO members against other WTO members with whom there are disagreements. The decisions require that the disagreeing parties must comply with international trading rules, thus resolving disputes. In December 2019 the United States blocked the appointment of new judges of the Appellate Body. If this issue is not resolved it is feared that as a result the WTO will be unable to continue to carry out this important function.

REAL WORLD FOCUS 15.6

US meat labelling provokes reaction in Canada and Mexico

In 2008, Canada and Mexico accused the United States of illegally requiring imported pork and beef to have country-of-origin labelling (COOL). According to officials, 'Country of origin labelling harms Canadian and Mexican livestock producers, as well as U.S. processors and producers.' Supporters of this labelling in the United States argued that it was important for consumers to know where the meat came from. Opponents argued that in view of food safety inspections already in place, the labelling did not contribute anything to food safety. Instead, they claimed that this was a form of hidden protection that created complex bureaucratic procedures resulting in higher prices of imported meat, making it less competitive in the US economy.

Figure 15.6: Wheaton, Maryland, USA. Meat labelling in the United States that violated WTO rules

Canada and Mexico took the case to the World Trade Organization, which found that COOL was in violation of international trade rules. In December 2015, the WTO announced that Canada was losing US$740 million annually and Mexico about US$228 million annually, as a result of COOL. In December 2015, the WTO therefore gave permission to both Canada and Mexico to retaliate by imposing tariffs on a range of imported products from the United States.

Reaction in the United States was swift. Faced with the prospect of retaliatory tariffs by Canada and Mexico, a US official said, 'If Canada and Mexico take steps to raise import duties on U.S. exports, it will only harm the economies of all three trading partners.'

In December 2015, the United States repealed the COOL scheme.

Sources: *www.cbc.ca; www.theglobeandmail.com; www.ictsd.org;*
CBS News 2015; Panetta 2015; International Centre for Trade and Sustainable Development 2016

Applying your skills

1 Identify the type of trade protection that COOL represented.

2 Using an *AD-AS* diagram, explain the statement that COOL harmed 'US processors and producers'. Using the same diagram, explain the effect that COOL had on consumers.

3 Using an international trade diagram, explain the US claim that if Canada and Mexico were to impose tariffs (import duties), '*it will only harm the economies of all three trading partners*'.

TEST YOUR UNDERSTANDING 15.4

1 What are the objectives of the World Trade Organization (WTO)?

2 What are some positive contributions of the WTO to the global trading system?

3 What are some criticisms of the WTO?

4 Explain some challenges faced by the WTO.

THEORY OF KNOWLEDGE 15.1

Is there a moral aspect in the economic argument in favour of free trade?

Perhaps the strongest economic argument in favour of free trade is that it leads to greater efficiencies in production and an improved allocation of resources. If the different types of trade protection discussed in this chapter were removed, there would be efficiency improvements. Yet everyone recognises that these efficiency gains involve both winners and losers: those who gain from trade protection lose from the removal of trade protection; and the losers from trade protection become the gainers after its removal.

In view of the fact that there are both gainers and losers, is it possible to argue in favour of free trade on a purely social scientific basis, without moral judgements? Welfare analysis shows that the gains to society are greater than the losses from a removal of trade barriers, but even so, there is still a moral judgement involved in the statement 'It's okay to sacrifice the well-being of some people for the sake of gains in the well-being of a larger number of people.' Economists (and others) recognise this point.

How then to get around this problem, so as to be able to recommend free trade on a scientific basis? Economists found the solution in what is known as the Hicks/Kaldor criterion[9] according to which if the winners can afford to compensate the losers and still be better off, then the removal of trade barriers and the switch to free trade is *scientifically* justified.

In practice, this could mean that the government taxes the gainers and then somehow pays the losers for their losses. However, in the real world this practically never happens. When it never happens, economists, for their part, can argue that they have done their job as economists, and from there on it is the governments' responsibility to pursue the right policies.

However, if economists recommend, and governments adopt free trade policies, *in full knowledge that there will not be any compensation of the losers*, is the recommendation still free of moral judgements? Or is it based on the same moral judgement as that noted above, that 'It's okay to sacrifice the well-being of some people for the sake of gains in the well-being of a larger number of people.'

We will return to the topic of free trade in Theory of knowledge 20.1 in Chapter 20.

Thinking points

- Do you agree that there is a moral judgement in the economic argument in favour of free trade?
- If there is a moral judgement, does it matter that it is generally 'covered up' and ignored in discussions of free trade?
- Do economists have a moral responsibility toward societies when making policy recommendations, or can they make recommendations in the belief that they are functioning purely as social scientists?

[9] This work is based on the work of British economist John Hicks and Hungarian-born economist Nicholas Kaldor. Hicks and Kaldor were among the more influential economists of the 20th century.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Examine whether the country you live in is a member of one or more trading blocs. Examine, too, whether it is currently in the process of negotiating the formation of or entry into a new trading bloc. In each case, what form of integration is involved: free trade agreement, customs union or common market? Are the agreements bilateral or regional? Choose one or more of these trading blocs and try to determine which stakeholders are likely to be better off or worse off as a result of the agreement. Try to identify the possible advantages and disadvantages for stakeholders and for the economy.

2 Since the time of writing of this book, the trade war that began in early 2018 between the United States and several other countries is likely to have evolved. Examine the current situation: was the trade war resolved with tariffs gradually being phased out; did it escalate with more tariffs being imposed; has it spread to more countries which have been affected by trade barriers? What, if any, has been the role of the World Trade Organization in resolving the disputes? Was or is your country involved in the trade war? If so what have been some of the consequences?

3 Examine the most recent developments regarding the WTO's role in resolving disputes. Have there been any new appointments to the Appellate Body? What are the prospects for this role of the WTO? What possible factors might be leading to the developments?

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 16

Exchange rates and the balance of payments

BEFORE YOU START

- Perhaps you have travelled to a foreign country or know others who have. When travelling, you pay for goods and services using that country's currency so you need to exchange your currency for the foreign currency. What do you think determines the 'price' you pay to exchange currencies?

This chapter examines the monetary or financial side of international links between countries. We will discover how exchange rates are determined and we will examine the positive or negative consequences of exchange rate changes. We will then study the balance of payments accounts, in order to see how countries record and oversee international money flows.

Currencies of different countries

16.1 Floating exchange rates

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain how exchange rates are determined by demand and supply of a currency in the foreign exchange market (AO2)
- draw diagrams illustrating exchange rate determination and exchange rate changes showing appreciation and depreciation (AO4)
- calculate prices of goods in different currencies (AO4)
- explain the various factors that may cause changes in demand and/or supply of a currency (AO2)
- calculate changes in the value of currencies, arising from changes in currency demand or supply (AO4)

Demand and supply of foreign exchange

At any moment in time, there is a continuous flow of money in and out of every country in the world. This happens because the residents of each country, whether individuals, or groups of individuals, or firms, or the government, have *transactions* (or dealings of any kind involving money) with the residents of other countries. International transactions involve the use of different national currencies, known as **foreign exchange**. These national currencies are traded for each other in the foreign exchange market, where individuals, firms, banks, other financial institutions and governments buy and sell currencies. The foreign exchange market is not a centralised meeting place, but involves any location where one currency can be exchanged for another, and any individual or organisation that engages in the exchange of one currency for another.

Suppose you are a resident of Russia travelling to Denmark. You will want to exchange some of your roubles for Danish kroner. To do this, you 'sell' your roubles and 'buy' Danish kroner in the foreign exchange market. The reason you must do so is that the residents of Denmark usually prefer to be paid in Danish kroner. Likewise, residents of European Monetary Union countries (the countries of the European Union that have adopted the euro, also known as 'euro zone' countries) want to be paid in euros; and residents of Chile want to be paid in Chilean pesos.

The foreign exchange market, like any market, is made up of demand and supply of currencies. As a traveller from Russia, when you change your roubles into Danish kroner, you *demand* Danish kroner, and you *supply* roubles in the foreign exchange market.

Similarly, suppose residents in the United Kingdom and Japan want to trade with each other. When Japanese residents import from the United Kingdom, they buy British pounds with which to pay UK exporters; they therefore *demand* British pounds in the foreign exchange market. To receive the pounds, they sell or *supply* yen in the foreign exchange market. When UK residents import from Japan, they *demand* yen, and *supply* pounds in the foreign exchange market to receive the yen.

This simple two-country example illustrates the equivalence between the demand for a foreign currency and the supply of a domestic currency. The demand for pounds is equivalent to a supply of yen, and the demand for yen is equivalent to a supply of pounds. There is a similar equivalence in the real world, where there are many different currencies: the demand for yen is equivalent to the supplies of all other currencies offered (or sold) in the foreign exchange market to buy yen. Similarly, the demand for pounds is equivalent to the supply of all other currencies offered to buy pounds.

The demand for foreign currencies generates a supply of domestic currency; and demand for the domestic currency generates a supply of foreign currencies. In a simple two-currency example using pounds and yen, it follows that:

demand for pounds $\Leftrightarrow$ supply of yen

demand for yen $\Leftrightarrow$ supply of pounds

Exchange rates

If national currencies can be exchanged for each other, there must be some mechanism of establishing the 'value' of each currency. This is done through the **exchange rate**, which is the value of one currency expressed in terms of another. Consider a *hypothetical* exchange rate between the US dollar and the euro:

- number of dollars per euro: 1.5 dollars = 1 euro
- number of euros per dollar: 0.67 euro = 1 dollar

The first expression gives the 'value' or 'price' of 1 euro in terms of dollars, showing how many dollars must be given up to buy 1 euro, as well as how many dollars can be gotten if one euro is given up. The second gives the 'value' or 'price' of 1 dollar in terms of euros, showing how many euros must be given up to buy 1 dollar, as well as how many euros can be gotten in exchange for one dollar. The two expressions are equivalent. They have to be, since the value of each currency is expressed in terms of the other.

To understand how exchange rates are determined, we will consider two 'pure' exchange rate systems: the *floating exchange rate system*, and the *fixed exchange rate system*. We will also study the actual system used by most countries today, known as a managed float, or *managed exchange rate system*, which lies in between the two 'pure' systems, though it is closer to floating exchange rates.

Determination of floating exchange rates

In a **floating exchange rate system** (also known as flexible exchange rate system), exchange rates are determined by market forces, or the forces of demand and supply, with no government or central bank intervention in the foreign exchange market.

Consider a simplified world with two currencies, the US dollar and the euro. In a floating system, the 'price' of the dollar and the 'price' of the euro are each determined in the same way that prices are determined in any free market. However, as we know from our discussion above, the 'price' of one currency is always expressed in terms of another currency, as there is no independent unit we can use to express the value of currencies.

Figure 16.1(a) shows the market for dollars. The horizontal axis measures the quantity of dollars, and the vertical axis measures the price of dollars in terms of euros. The demand curve represents the demand for dollars, and the supply curve represents the supply of dollars.

The demand for dollars, shown by a familiar downward-sloping curve, comes from euro zone residents who need dollars to carry out transactions in the United States: euro zone importers buying goods from the US, euro zone investors who want to invest in the US, consumers going on holiday to the US, etc. The downward slope of the curve indicates that as the price of dollars in terms

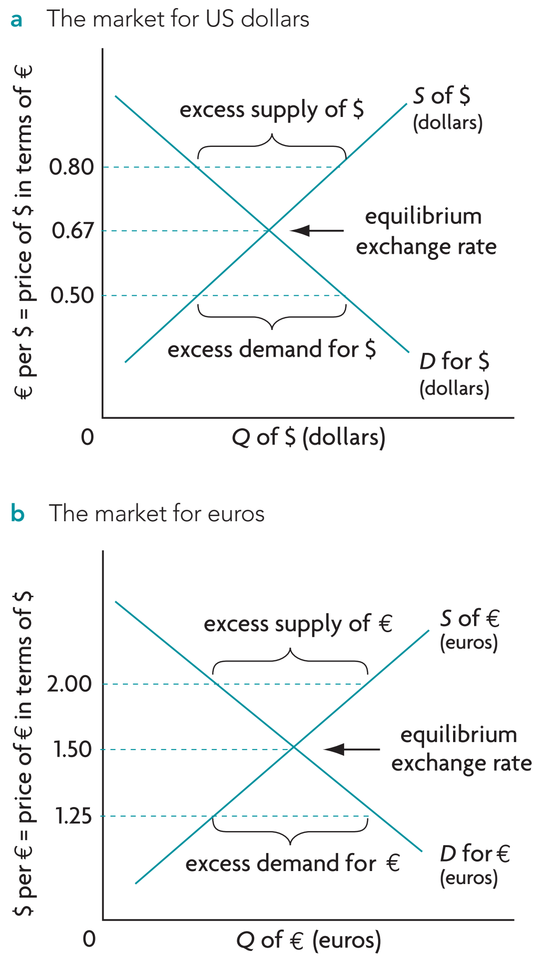


Figure 16.1: Exchange rate determination in a floating exchange rate system

of euros increases, euro zone residents buy fewer dollars. For example, if 0.80 euro are needed to buy 1 dollar, euro zone residents buy fewer dollars than if 0.5 euro are needed to buy 1 dollar.

The supply of dollars, shown by a familiar upward-sloping curve, comes from US residents who would like to 'sell' dollars to buy euros: US residents who would like to import goods from euro zone countries, or who want to take a holiday in a euro zone country, or who plan to invest in a euro zone country, etc. To see why the supply-of-dollars curve is upward sloping, consider that if the price of dollars is 0.5 euros per dollar, US residents need to supply 1 dollar to buy 0.5 euro worth of euro zone goods; if the price of dollars increases to 0.8 euro per dollar, then by selling 1 dollar US residents can buy 0.8 euro worth of euro zone goods. As the price of dollars goes up, euro zone goods become cheaper, and so more dollars are supplied.

Therefore, as the price of dollars in terms of euros increases, the quantity of dollars supplied increases.

The intersection of the demand and supply curves determines the equilibrium 'price' of the dollar in terms of the euro; this is the equilibrium exchange rate, which is 0.67 euro per dollar. If the exchange rate were higher than the equilibrium, say at 0.8 euro per dollar, there would be an excess supply of dollars. At any exchange rate lower than the equilibrium, such as 0.5 euro per dollar, there would be an excess demand for dollars.

> In a floating exchange rate system, the equilibrium exchange rate is determined by the forces of demand and supply at the point where the quantity of a currency demanded equals quantity supplied, without any government or central bank intervention.

Figure 16.1(b) shows the market for euros. Since the demand for dollars is equivalent to the supply of euros, and the supply of dollars is equivalent to the demand for euros, it follows that when we determine the 'price' of dollars, we also determine the 'price' of euros. The demand curve, showing demand for euros by US residents who want to buy euros to import, travel, invest, etc. in euro zone countries, is a reflection of the supply-of-dollars curve in part (a). The supply curve, showing the supply of euros from euro zone residents who want to buy dollars to import, travel, invest, etc. in the United States, is a reflection of the demand-for-dollars curve in part (a). The intersection of the demand-for-euros and supply-of-euros curves determines the equilibrium exchange rate, or the 'price' of the euro in terms of the dollar, which is 1 euro = 1.5 dollars.

The two equilibrium exchange rates in Figure 16.1 are equivalent to each other. At any other exchange rate, the markets are in disequilibrium. When the price of dollars in terms of euros is 0.80, there is an excess supply of dollars that corresponds to an excess demand for euros. When the price of dollars in terms of euros is 0.50, the excess demand for dollars reflects an excess supply of euros.

Exchange rate changes: appreciation and depreciation

Once an exchange rate settles at its equilibrium value, it will remain there until there is a change in demand or supply of the currency, expressed as a shift in the currency demand or supply curve.

Figure 16.2 shows how a currency changes in value in response to changes in demand or supply of the currency. An increase in the value of a currency (such as the dollar) in a floating exchange rate system is called an **appreciation**. This is shown in Figure 16.2(a) to result from an increase in the demand for dollars, causing the demand-for-dollars curve to shift to the right, resulting in a higher exchange rate e_2. In addition, it may result from a decrease in the supply of dollars, shown in part (b), causing the supply-of-dollars curve to shift to the left.

A fall in the value of a currency (the dollar) in a floating exchange rate system is called a **depreciation**. The value of the dollar decreases in Figure 16.2(c) due to a fall in the demand for dollars, causing a leftward shift in the demand-for-dollars curve, hence a fall in the value of the dollar to e_2; or in part (d) due to an increase in the supply of dollars, causing a rightward shift in the supply-of-dollars curve.

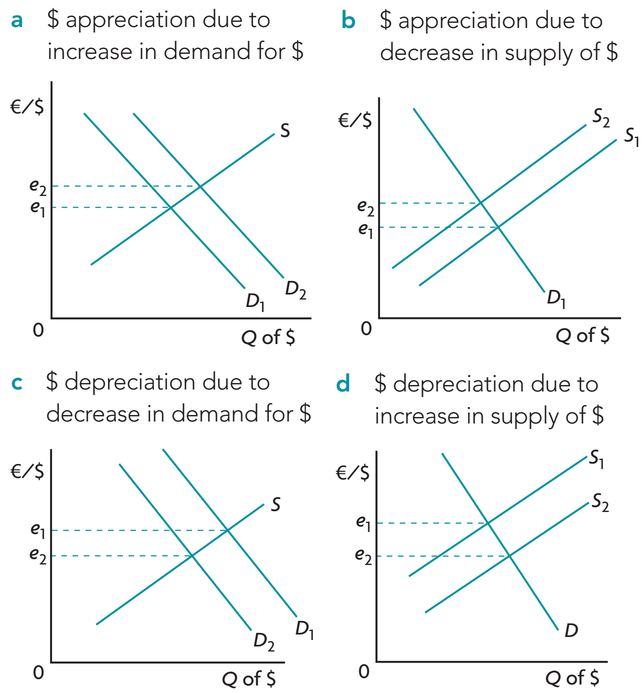


Figure 16.2: Exchange rate changes in a floating exchange rate system

The same general principles apply to the international system consisting of many national currencies. When one currency appreciates it does so against all others in a floating exchange rate system, meaning that all others depreciate relative to it. When a currency depreciates, all other currencies appreciate relative to it.

In a floating exchange rate system, appreciation (increase in value) and depreciation (decrease in value) of a currency occur as a result of changes in demand or supply for a currency in the absence of any government or central bank intervention.

TEST YOUR UNDERSTANDING 16.1

1 State why exchange rates are measured in terms of the quantity of another currency.

2 Explain how exchange rates are determined in a floating exchange rate system.

3 Explain how each of the following affects the demand or supply of the US dollar:

   a An investor from the United States invests in the stock market of Brazil.

   b A Peruvian importer buys goods from the United States.

   c A US resident travels to China as a tourist.

   d The United States contributes to foreign aid in Africa.

4 Define appreciation and depreciation of a currency.

5 Suppose the United States would like to increase its imports from euro zone countries. Using diagrams, show the effects on

   a the value of the US dollar, and

   b the euro.

   c State which currency will appreciate and which will depreciate.

Causes of changes in exchange rates

In the real world, there are numerous ongoing changes in demand and supply of currencies, causing exchange rates of most currencies around the world to fluctuate on a daily or even hourly basis.

Changes in currency demand

Currency demand involves factors that lead to *inflows of funds* into a country.

Exports and factors affecting exports

Foreign demand for exports of goods. Changes in foreign demand for a country's exports affect its exchange rate. If there is an increase in foreigners' demand for Swiss watches, the demand for Swiss francs increases, the demand-for-francs curve shifts to the right, and the franc appreciates. If foreigners' demand for Swiss watches falls, the demand for francs decreases and the franc depreciates.

Foreign demand for exports of services, such as tourism services. An increase in tourism from abroad represents an increase in exports of services (tourism services). Tourists coming from abroad demand more of the country's currency and the currency appreciates. If tourism from abroad falls, demand for the currency falls and the currency depreciates.

The rate of inflation relative to other countries. If Sweden experiences a lower rate of inflation than other countries, demand for its exports increases as other countries now find them relatively less expensive. The increase in Swedish exports causes demand for Swedish kronor (the Swedish currency) to shift to the right causing the Swedish kronor to appreciate. If Sweden's rate of inflation is higher than that of other countries, foreign demand for its exports falls, and demand for the Swedish kronor falls, and the kronor depreciates.

Relative growth rates. If Kenya's trading partners experience high economic growth, this means that their incomes are rising and so they demand more exports from Kenya, the demand for the Kenyan shilling increases, the demand-for-shilling curve shifts to the right, and the shilling appreciates. By contrast, lower economic growth or recession in Kenya's trading partners will lead to a depreciation of the shilling.

Investment and factors affecting investment

Inward foreign direct investment (FDI) and portfolio investment. There are two types of investment by foreigners: **foreign direct investment** (investment by multinational corporations in productive facilities; see Chapter 20) and **portfolio investment** (financial investments, such as purchase of stocks and bonds).

Both types of investment have the same impact on exchange rates. If the investment is *inward* (coming into the country), foreigners are bringing in funds from abroad by demanding the domestic currency. If foreigners want to increase their investments in China, they must buy Chinese yuan, and the demand-for-yuan curve shifts to the right, appreciating the yuan. If foreigners decrease their investment in China, the demand for yuan falls and the yuan depreciates. This applies to both FDI and portfolio investment.

Relative interest rates. Financial capital (defined in Chapter 1, Section 1.1) refers to funds that are used to make financial investments, or investments that receive a return based partly on the rate of interest. (Financial investments are a type of *portfolio investment*, defined above.) For example, savings deposits in banks and purchases of bonds depend on the rate of interest that can be received. The higher the rate of interest in a country, the more attractive are the savings deposits and bonds. If interest rates in the United Kingdom increase relative to interest rates in other countries, financial investments become more attractive, more financial capital will flow in, demand for British pounds increases, the demand-for-pounds curve shifts to the right, and the pound appreciates. Similarly, if interest rates in the United Kingdom fall relative to interest rates in other countries, less financial capital will flow in and the pound will depreciate. Very often *expectations* of financial investors that interest rates will rise or fall in a country are enough to cause currency appreciation or depreciation.

Other factors affecting currency demand

Inward flow of remittances. **Remittances** involve a transfer of money from one country to another, in most cases by foreign workers who send money from their earnings in the country of residence to their family in their home country. In 2018, India and China were by far the largest receivers of remittances of their workers abroad who sent money home. An increase in remittances into a country sent from abroad, for example into India, leads to an increase in the demand for the Indian rupee and a rupee appreciation. A fall in remittances into India leads to a fall in demand for the Indian rupee hence a rupee depreciation.

Speculation that a currency will appreciate. Currency **speculation** involves buying and selling currencies to make a profit from changes in exchange rates. Buying and selling is based on expectations of future exchange rate changes. If currency speculators expect a country's currency to appreciate, they buy it in the hope of selling it later after its appreciation, thereby making a profit. However, as they buy the currency they may cause it to appreciate; there is therefore a self-fulfilling prophecy at work. Speculators can therefore cause exchange rate changes through their actions.

Central bank intervention to increase the value of a currency. Every central bank holds reserves of foreign currencies that it sometimes buys or sells to influence the value of the domestic currency (see below). If the central bank wants to increase the value of the domestic currency, it demands (buys) the domestic currency (by selling foreign currencies), the demand curve shifts to the right, and the currency appreciates.

Changes in currency supply

Currency supply involves factors that lead to *outflows of funds* from a country.

Imports and factors affecting imports

Domestic demand for imports of goods. Changes in a country's demand for imports affect its exchange rate. If consumers in the United States import more foreign-made cars, US importers must buy more foreign currencies, and so they supply (sell) more US dollars in the foreign exchange market. As the supply of dollars increases, the supply-of-dollar curve shifts to the right, and the dollar depreciates. If the US demand for foreign cars falls, there is a leftward shift in the supply-of-dollars curve and the dollar appreciates.

Domestic demand for imports of services, such as tourism. Imports of tourism services refers to tourists travelling abroad to foreign countries. To do so they sell domestic currency to buy foreign exchange that they can use abroad. Increasing numbers of French tourists travelling abroad outside of the euro zone means the supply of euro will increase, and so the euro will depreciate. On the other hand if the number of French tourists travelling abroad outside of the euro zone falls, this means the supply of the euro will decrease so the euro will appreciate.

The rate of inflation relative to other countries. If Sweden's rate of inflation is lower than that of other countries, demand for imports decreases as these are relatively more expensive compared to domestically produced goods, and supply for the Swedish kronor decreases, causing it to appreciate (note this reinforces the effect of greater demand for exports discussed above). If Sweden experiences a higher rate of inflation compared to other countries, imports from other countries with lower inflation rates will increase as Swedes find them cheaper. There is a rightward shift in supply of kroner, causing the Swedish kronor to depreciate. (This reinforces the effect of lower demand for exports noted earlier.)

Relative growth rates. If Kenya experiences higher economic growth with growth in incomes its demand for imports will rise, therefore the supply of Kenyan shilling increases and the shilling depreciates. On the other hand lower growth rates in Kenya mean lower income growth, the demand for imports will decrease, therefore the supply of Kenyan shilling decreases and the shilling appreciates.

Investment and factors affecting investment

Outward foreign direct investment (FDI) and portfolio investment. Investment by multinational corporations in productive facilities in countries outside of the home country, as well as portfolio investments (such as purchases of stocks and bonds) outside the home country, involve funds flowing out thus giving rise to supply of the domestic currency. If Thai investors increase their investments outside of Thailand there will be an increase in the supply of Thai baht and the baht will depreciate. If Thai investors decrease their investments abroad, the supply of baht will decrease and the baht will appreciate.

Relative interest rates. We have seen that interest rates influence the flow of funds in search of higher rates of return from country to country. As a result, interest rates and exchange rates tend to move together, *ceteris paribus*. If interest rates in the United Kingdom fall relative to interest rates in other countries, more financial capital will flow out of the country, and the supply of pounds will increase, causing depreciation. But if interest rates in the United Kingdom increase relative to interest rates in other countries, less financial capital will flow out of the United Kingdom, because domestic investors will want to take advantage of the higher domestic interest rates, the supply of British pounds decreases, and the pound appreciates. (Note that these effects reinforce those resulting from changes in demand for a currency.)

Other factors affecting currency supply

Outward flow of remittances. These are outward flows of money by foreign workers in a country who send money home. In 2018, the United States was by far the leading country in the world of remittance outflows. The most important countries that received remittances from the United States were Mexico, China, India and the Philippines. An increase in remittances sent home by workers living in the United States result in an increase in the supply of US \$, so the \$ depreciates. A decrease in remittances sent home by workers in the United States leads to a fall in the supply of US \$ and \$ appreciation.

Speculation that a currency will depreciate. If currency speculators expect a country's currency to depreciate, they sell it in the hope of buying it later after its depreciation, thereby making a profit. However, as they sell the currency they may cause the supply curve to shift to the right making the currency depreciate; there is therefore a self-fulfilling prophecy at work.

Central bank intervention to decrease the value of a currency. If the central bank wants to lower the value of the domestic currency, it supplies (sells) more of the domestic currency (by buying foreign currencies), the supply of the currency increases and the currency depreciates.

When both currency demand and supply change at the same time

We have seen that some events cause both demand and supply *of the same currency* to change at the same time:

Rate of inflation relative to other countries. If Sweden has a lower rate of inflation relative to other countries, demand for its exports increases and demand for imports decreases; both these factors lead to currency appreciation.

Interest rates relative to other countries. An increase in interest rates in the United Kingdom attracts financial capital, causing the demand for the British pound to increase and supply of the British pound to decrease; both these factors lead to pound appreciation.

Factors that affect currency demand Inflows of funds		Factors that affect currency supply Outflows of funds	
Increase in currency demand leads to appreciation Figure 16.2(a) D curve shifts right	**Decrease in currency demand leads to depreciation Figure 16.2(c) D curve shifts left**	**Decrease in currency supply leads to appreciation Figure 16.2(b) S curve shifts left**	**Increase in currency supply leads to depreciation Figure 16.2(d) S curve shifts right**
Increase in foreign demand for exports of goods and services	Decrease in foreign demand for exports of goods and services	Decrease in domestic demand for imports of goods and services	Increase in domestic demand for imports of goods and services
Lower inflation leading to increase in foreign demand for exports	Higher inflation leading to decrease in foreign demand for exports	Lower inflation leading to decrease in domestic demand for imports	Higher inflation leading to increase in domestic demand for imports
High growth rates of trading partners leading to increase in foreign demand for exports	Low growth rates of trading partners leading to decrease in foreign demand for exports	Low domestic growth rate leading to decrease in domestic demand for imports	High domestic growth rate leading to increase in domestic demand for imports
Increase in inward investment	Decrease in inward investment	Decrease in outward investment	Increase in outward investment
Higher interest rates leading to more inward financial investment	Lower interest rates leading to less inward financial investment	Higher interest rates leading to less financial investment by domestic residents in foreign countries	Lower interest rates leading to more financial investment by domestic residents in foreign countries
Increase in inflow of remittances	Decrease in inflow of remittances	Decrease in outflow of remittances	Increase in outflow of remittances
Speculators expect currency X will rise so they buy currency X	-	-	Speculators expect currency X will fall so they sell currency X
Central bank buys the domestic currency	-	-	Central bank sells the domestic currency

Table 16.1: Summary of causes of exchange rates changes

TEST YOUR UNDERSTANDING 16.2

1 For each of the following events, draw exchange rate diagrams and show how a shifting currency demand or supply curve causes an appreciation or depreciation of the currency:

a An increase in interest rates in the United States relative to the rest of the world; show the impact on the US dollar and on the British pound.

b Philippines has reduced remittances from Canada; show the impact on the Philippine peso and the Canadian dollar.

c An increase in the rate of inflation in Thailand relative to its trading partners; show the impact on the baht (Thailand's currency) and on the ringgit (Malaysia's currency; Malaysia is a trading partner of Thailand).

CONTINUED

d Currency speculators believe that the Brazilian real will appreciate; show the impact on the Brazilian real.

e Japan is in recession, and incomes are falling; show the impact on the Japanese yen.

f China experiences an increase in tourism; show the impact on the Chinese yuan.

g Fashion favours Indian textiles; show the impact on the Indian rupee.

h Foreign direct investment from Mauritius into India increases; show the effect on the Mauritian rupee and the Indian rupee.

i The central bank of Botswana does not want to allow the pula to depreciate; show the impact on the pula.

2 Some of the items appearing in Table 16.1 give rise to changes in both demand and supply *of the same currency*. Identify what these are and explain why both currency demand and supply are affected.

Calculations using exchange rates

Calculating the value of a currency in terms of another

Example 1

A hypothetical exchange rate of 1.5 dollars = 1 euro gives us the price of 1 euro in terms of dollars. If we want to find the price of 1 dollar in terms of euros, we divide the unit currency (euro) by the other currency (dollars). Therefore,

$$1 \text{ dollar} = \frac{1}{1.5} \text{ euro} = 0.67 \text{ euro}$$

The expressions 1.5 dollars = 1 euro, and 0.67 euro = 1 dollar are equivalent.

Example 2

The hypothetical exchange rate 0.37 Russian rouble (RUB) = 1 Japanese yen (JPY) gives the price of 1 yen in terms of roubles. Find the price of 1 rouble in terms of yen.

$$1 \text{ rouble} = \frac{1}{0.37} \text{ yen} = 2.70 \text{ yen}$$

The expressions 0.37 rouble = 1 yen, and 2.70 yen = 1 rouble are equivalent.

In the real world, exchange rates are usually expressed to many decimal places. For example, we may find that 1 ruble = 2.70135 yen. Even a very small change in an exchange rate can amount to large differences in the total values being traded if large quantities of money are involved.

Calculating prices in different currencies

Suppose an importer in the United Kingdom imports wine from France (which is a euro zone country). The hypothetical exchange rate between British pounds (£) and euros (€) is £1.22 = €1. The importer wants to import 1000 bottles at the price of €5 per bottle. Since the importer will supply £ to make the payment in €, she is interested in finding the cost in £.

In terms of euros, the cost is 1000 × €5 = €5000. To find this amount in pounds, we simply multiply it by 1.22 (since €1 = £1.22), and we find 1.22 × 5000 = £6100.

Calculating changes in the value of a currency from a set of exchange rate data

Interpreting exchange rate data

Suppose you are given a set of data on the following exchange rate changes over time. Here we consider an imaginary country called Bopland whose national currency is the bople.

January 2020	1.22	July 2020	1.40
February 2020	1.25	August 2020	1.37
March 2020	1.33	September 2020	1.45
April 2020	1.39	October 2020	1.58
May 2020	1.47	November 2020	1.63
June 2020	1.43	December 2020	1.69

Table 16.2: US$ per 1 bople; average monthly exchange rates

The data in Table 16.2 show the value of 1 bople in terms of US \$. Has the bople appreciated or depreciated in the period from January to December 2020? In January, 1 bople was worth \$1.22, while in December it was worth \$1.69. The value of the bople increased, in other words the bople appreciated relative to the \$. However, it did not appreciate every month. In June, July and August it depreciated (or lost some value) compared to the previous month.

Calculating percentage changes in the value of a currency

What was the percentage change in the value of the bople between January and December 2020? (For a review of percentage changes, see 'Quantitative techniques' chapter in the digital coursebook.)

% change in the bople (January–December) =

$$\frac{1.69-1.22}{1.22}\times 100=\frac{0.47}{1.22}\times 100=38.52\%$$

Therefore the bople appreciated by 38.52% relative to the \$ during 2020.

The bople appreciation relative to the \$ corresponds to a \$ depreciation relative to the bople. To find the percentage change in the value of the \$ for this period, we must first find the 'price' of the \$ in terms of boples. In January, \$1.22 = 1 bople, therefore:

$$\$1=\frac{1}{1.22\text{ bople}}=0.82\text{ bople}$$

In December, \$1.69 = 1 bople, therefore:

$$\$1=\frac{1}{1.69\text{ bople}}=0.59\text{ bople}$$

We can now use this information to find the percentage change in the \$:

% change in the \$ (January–December) =

$$\frac{0.59-0.82}{0.82}\times 100=\frac{-0.23}{0.82}\times 100=-28.05\%$$

The negative percentage change indicates a fall in the value of the \$; therefore the \$ depreciated by 28.05% relative to the bople in 2020.

This exercise indicates that although an appreciation of currency X relative to currency Y is equivalent to a depreciation of currency Y relative to currency X, the percentage changes *are not the same* (with one being positive and the other negative).[1]

TEST YOUR UNDERSTANDING 16.3

1 a 1 Canadian dollar = 0.99 US dollar. Calculate the value of 1 US dollar in terms of Canadian dollars.

b 1 Indian rupee = 1.84 Japanese yen. Calculate the 'price' of 1 Japanese yen in terms of Indian rupees.

c 1 Japanese yen = 1.34 Sri Lanka rupee. Calculate the value of 1 Sri Lanka rupee in terms of Japanese yen.

d 1 British pound= 1.62 Canadian dollars. Calculate the value of 1 Canadian dollar in terms of British pounds.

2 The price of item X in India is 50 Indian rupees. Using the exchange rates in question 1, find its price in

a Japanese yen, and

b Sri Lanka rupees.

c Importers in Japan and Sri Lanka want to import 1000 units of item X. Calculate their price in yen and rupees, respectively.

3 The price of item Y in Canada is 75 Canadian dollars. Using the exchange rates in question 1, find its price in

a US dollars, and

b British pounds.

c What is the cost of 5000 units of item Y in dollars and pounds?

[1] The reason for this is that percentage changes are calculated relative to an initial value. Since the initial values are different for the two currencies, their percentage changes are also different.

CONTINUED

4 On 1 June 2010, 1 British pound was worth US \$1.46; on 1 November 1 British pound was worth US \$1.60.

a Identify which currency appreciated and which depreciated.

b Calculate the percentage appreciation of the appreciating currency.

c Calculate the percentage depreciation of the depreciating currency.

5 a Using the table below, determine which currency appreciated, and which depreciated from the 1st to the 30th of September.

b Calculate the percentage appreciation of the appreciating currency.

c Calculate the percentage depreciation of the depreciating currency.

Euros to 1 US\$ (value of 1 US dollar per euro)

1 Sept. 2010	1.2800
8 Sept. 2010	1.2697
16 Sept. 2010	1.2989
23 Sept. 2010	1.3364
30 Sept. 2010	1.3611

16.2 Consequences of changes in exchange rates: an evaluation

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- evaluate the consequences of changes in exchange rates on economic variables such as inflation, unemployment, economic growth, the current account balance, standards of living (AO3)
- draw *AD-AS* diagrams to show consequences of exchange rate changes on important macroeconomic variables (AO4)

When a currency appreciates, a unit of it can buy more of other currencies, therefore it can buy more foreign goods; foreign goods, or imports, become cheaper. As a result we expect imports to increase. At the same time, foreigners need to give up more of their currencies to buy the domestic currency, so that foreigners can buy fewer of the domestic goods; therefore domestic goods, or exports, become more expensive to foreigners and we expect that exports will fall. This means that net exports or $(X - M)$ will decrease.

When a currency depreciates, the opposite happens. It loses its value relative to other currencies, so imports become more expensive while exports become cheaper to foreigners. The depreciated currency can therefore buy fewer foreign goods, or imports, while foreigners can buy more domestic goods, or exports. This means that net exports, or $(X - M)$ will increase. (HL students will learn that this depends on the Marshall-Lerner condition; see Chapter 17.)

Currency appreciation can be expected to result in a decrease in net exports $(X - M)$ while currency depreciation can be expected to lead to an increase in net exports $(X - M)$.

Many of the consequences of exchange rate changes follow from the above results.

Effects on the rate of inflation

Exchange rate changes can affect inflation in two ways.

Demand-pull inflation

Exchange rate changes affect aggregate demand by influencing net exports $(X - M)$. A currency depreciation, by making exports cheaper and imports more expensive, works to increase the quantity of exports and lower the quantity of imports, thus increasing net exports $(X - M)$. An increase in net exports results in a rightward shift of the aggregate demand curve. Using the Keynesian *AD-AS* model, we can see that whether or not this will cause demand-pull inflation depends on where the economy is in the business cycle. If it is in recession, an increase in AD will not cause demand-pull inflation, however if the economy is producing at or close to potential output, inflationary pressures will result due to excess aggregate demand (see Section 10.2, Chapter 10).

A currency appreciation will work to reduce demand-pull inflationary pressures in an economy due to a decrease in net exports ($X - M$), causing AD to fall.

Cost-push inflation

A currency depreciation, as we have seen, makes imports more expensive. If domestic producers are heavily dependent on imported factors of production, their costs of production increase, resulting in a leftward shift of the $SRAS$ curve resulting in cost-push inflation (see Section 10.2, Chapter 10).

A currency appreciation, by making imports less expensive, results in a rightward shift of the $SRAS$ curve, lowering inflationary pressures in the economy (lower cost-push inflation).

Effects on economic growth

The effects of exchange rate changes on economic growth work directly through net exports and aggregate demand, discussed above, but they may also have effects on aggregate supply.

A currency depreciation increases net exports, increasing aggregate demand, thus causing an increase in real GDP produced. This is short-term growth (see Chapter 11). Also, if the growth of export industries leads to increased investment spending in the domestic economy (production of capital goods), there may be effects on aggregate supply, causing increases in potential output (rightward shifts in the $LRAS$ or Keynesian AS curves). This is long-term growth.

However, on the negative side, to the extent that there is cost-push inflation (see above) there may a downward pressure on real GDP which may fall due to the decrease in short-run aggregate supply (leftward shift in the $SRAS$ curve).

What will happen to real GDP depends on which of the two effects is stronger: the upward effect due to the increase in aggregate demand or the downward effect due to the decrease in short run aggregate supply. On the other hand, a currency appreciation, by reducing net exports is likely to reduce growth of real GDP.

Effects on unemployment

We have seen that a currency depreciation increases net exports and therefore aggregate demand. This causes a fall in cyclical unemployment if the economy is in a recessionary gap. If the economy is at or close to potential GDP, the increase in aggregate demand may cause a *temporary* decrease in natural unemployment, however this will come with strong demand-pull inflationary pressures.

Employment in export industries is likely to increase since exports are likely to rise. In addition employment in industries producing goods that compete with imports is also likely to increase since imports are expected to fall with depreciation.

On the other hand, as we have seen above, depreciation may lead to cost-push inflation. This involves lower real GDP due to the decrease in short run aggregate supply (leftward shift of the $SRAS$ curve), hence an increase in unemployment. The overall effect on unemployment depends on which of the two effects is stronger: the upward effect due to the increase in aggregate demand or the downward effect due to the decrease in short-run aggregate supply.

A currency appreciation, by reducing net exports and aggregate demand, will create a recessionary gap and therefore lead to cyclical unemployment if the economy begins at or close to potential output, or will lead to an increase in cyclical unemployment if it is already in a recessionary gap.

Employment in export industries as well as in industries producing goods that compete with imports is likely to decrease since exports are expected to fall while imports are expected to increase.

Effects on the current account balance

The *current account balance* will be studied in Section 16.4. For now we can say that it consists mainly of the 'balance' of exports and imports of goods and services, specifically the value of exports minus the value of imports, or more simply $X - M$. For reasons we will discover later in this chapter, major differences between the value of exports and the value of imports are considered undesirable, especially if they persist over long periods.

As we know, depreciation is likely to cause imports to decrease and exports to increase. If a country has an excess of imports over exports (a trade 'deficit'), its deficit is likely to become smaller after a period of time. If it has excess of exports over imports to begin with (a trade 'surplus'), its trade surplus is likely to become larger. An appreciation, by contrast, will cause imports to increase and exports to fall, thus having the opposite effects on the current account balance. These points will become clearer to you after you have studied Section 16.4 below. (HL students will study this topic further under the Marshall-Lerner condition in Chapter 17.)

Effects on foreign debt

A depreciation, by lowering the value of the domestic currency, causes the value of foreign debt to increase. Suppose Mountainland owes foreigners $1000, and initially has an exchange rate of Mnl 1.5 = $1 (Mnl is Mountainland's national currency); its foreign debt is therefore Mnl 1500. If the Mnl depreciates, so that now Mnl 2 = $1, Mountainland's foreign debt of $1000 becomes Mnl 2000. This is a problem faced by many developing countries, which find themselves having a larger debt burden if their currency depreciates (see Chapter 20). On the other hand, a currency appreciation causes the value of foreign debt to fall.

Effects on living standards

When a currency depreciates, it causes imported goods to become more expensive in the domestic economy, therefore residents become worse off as all imported goods become more expensive. If the country depends on imports of oil, residents will be affected by increased prices of gasoline as well as heating oil. In addition, if there are important imported inputs in production, such as oil or capital goods, there will be cost-push inflation, that will increase the general price level, making the cost of living higher and therefore reducing the real incomes of residents. The upward effects on aggregate demand (see above) may add to the inflationary pressures through demand-pull inflation. Further, travellers abroad such as tourists to other countries will find their holidays have been made more expensive by the depreciation. The effects on unemployment are likely to be mixed, as explained above in the section on unemployment.

If, on the other hand, a currency appreciates, the effects on living standards of the residents are likely to be positive. Prices of imported goods fall, imported inputs become less expensive, leading to a downward pressure on the rate of inflation, which is reinforced by the likely decrease in demand-pull inflation. Therefore, real incomes will increase. Travellers abroad will also benefit as the cost of travelling outside of the country will decrease. As regards unemployment, the effects here too are likely to be mixed.

TEST YOUR UNDERSTANDING 16.4

1 Using *AD-AS* diagrams, explain the likely effects on
   - a inflation in Canada if the Canadian $ appreciates,
   - b living standards in Belarus if the Belarusian ruble appreciates,
   - c unemployment in the United Kingdom if the British pound depreciates,
   - d inflation in Chile if the Chilean peso depreciates,
   - e living standards in Singapore if the Singapore $ depreciates,
   - f economic growth in Iceland if the Icelandic króna depreciates,
   - g unemployment in Sri Lanka if the Sri Lankan rupee appreciates, and
   - h economic growth in Morocco if the Moroccan dirham appreciates.

2 Explain the likely effects of
   - a currency appreciation, and
   - b currency depreciation on
     - i current account balance, and
     - ii foreign debt.

REAL WORLD FOCUS 16.1

The falling Indian rupee

The year 2018 was a difficult one for the Indian rupee, which fell significantly relative to the US dollar. Among the factors responsible for rupee weakness is the tightened monetary policy of the US Federal Reserve (the central bank of the United States), which raised interest rates four times in the course of the year. Higher oil prices contributed to a widening current account deficit in India, which depends heavily on imported oil. India imports as much as 70% of the crude oil it consumes. It is expected that in 2018 the current account deficit will widen to 2.8% of GDP compared to 1.9% the previous year.

Figure 16.3: Mumbai, India. Reserve Bank of India, the central bank

The rupee depreciation has contributed to inflationary pressures in India. The prices of consumer goods, such as soaps, detergents and shampoos, whose production depends on crude oil, will be affected. In addition petrol and diesel prices will rise.

Other goods that India imports in large quantities, such as fertilisers, medicines, iron ore and some food products, are also becoming more expensive due to the weak rupee.

Employment in industries that depend heavily on imports will also be negatively affected. For example, electronic consumer goods such as computers, television and mobile phones have imported components. Tourism will be influenced on account of higher airfares arising from the higher cost of fuel, as well as the cost of spending abroad which will increase in terms of rupees. The automobile sector will suffer due to the use of imported components, as well as loans that have to be repaid in foreign currencies.

Sources: *Ahona Sengupta, 'As Rupee Slumps to Record Low of 71, Here's Why It's Falling Like Never Before', News18, 31 August 2018*
Chitranjan Kumar, 'Five factors that affected Indian currency in 2018', Business Today, 23 December 2018

Applying your skills

1 Using an exchange rate diagram and evidence from the text, explain the effects on the Indian rupee of higher interest rates in the United States.

2 Using *AD-AS* diagrams, explain the effects of rupee depreciation on

  a demand-pull inflation, and

  b cost-push inflation.

3 Discuss the likely effects of rupee depreciation on growth, unemployment, the current account balance, foreign debt and standards of living.

4 Investigate how the value of the rupee has changed since 2018. Determine whether the rupee has depreciated further, or if it has appreciated, and examine the factors behind the changes.

16.3 Government intervention

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain exchange rate determination in a fixed exchange rate system, including currency devaluation and revaluation (AO2)
- draw a diagram to show how fixed exchange rates are maintained (AO4)
- explain exchange rate determination in a managed exchange rate system (AO2)
- explain overvalued and undervalued currencies (AO2)
- draw a diagram to show how equilibrium exchange rates are determined and change in a managed exchange rate system (AO4)
- evaluate fixed versus floating exchange rate systems (HL only) (AO3) (to be considered in Chapter 17)

Fixed exchange rates

Understanding a fixed exchange rate system

In a **fixed exchange rate system**, exchange rates are fixed by the central bank of each country at a particular level (or narrow range), and are not permitted to change freely in response to changes in currency supply and demand. The exchange rate is still determined by currency demand and supply, but these are manipulated by the central bank or government in order to arrive at the particular equilibrium that will give rise to the desired exchange rate. Therefore maintaining the value of a currency at its fixed rate requires constant intervention by the central bank or government. This intervention takes the form of buying and selling reserve currencies by the central bank, as well as making other adjustments in the domestic economy, all intended to shift the currency demand or supply curves in order to arrive at the predetermined equilibrium.

To see how this works, again consider Bopland and its national currency, the bople. The market for boples is shown in Figure 16.4(a). The central bank of Bopland has fixed the bople–US dollar exchange rate at 2 US dollars = 1 bople. Initially there is equilibrium in the bople market, at point A. Suppose there occurs a leftward shift in the demand for boples (because, for example, of a fall in demand for Bopland's exports), so the demand-for-boples curve shifts from D_1 to D_2. At the fixed exchange rate of 2 dollars = 1 bople, there is an excess supply of boples (the distance A–B). Under a floating exchange rates, the exchange rate would fall to 1.50 dollars per bople (point C), eliminating the excess supply; but if the fixed exchange rate of 2 dollars per bople is to be maintained, the central bank or government must intervene.

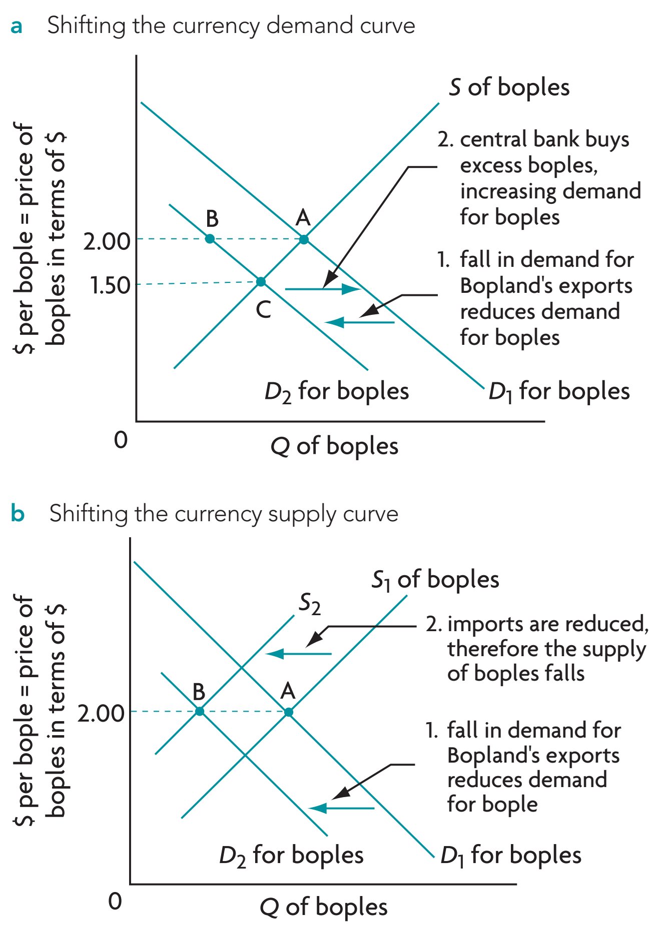


Figure 16.4: Fixed exchange rates: maintaining the value of the bople at 1 bople = $2.00

Intervention to maintain fixed exchange rates

Using official reserves to maintain the exchange rate

In the example above, there is an excess supply of boples, so the central bank of Bopland can intervene by buying the excess boples by selling some of its foreign currency reserves. This shifts the demand for boples curve back to D_1, and the fixed exchange rate of 2.00 dollars = 1 bople is maintained.

If there had been an excess demand for boples, the central bank would sell some boples by buying dollars.

However, if Bopland faces an excess supply of boples and a downward pressure on the bople's value over a long time, its central bank will eventually run out of foreign currency reserves, and will be unable to go on buying excess boples. At that point, the central bank or government must use additional methods to maintain the fixed exchange rate. However, each of these measures comes with some disadvantages.

The need for additional measures to maintain a fixed exchange rate arises primarily when there is an excess supply of the domestic currency over a long time. In the opposite situation, where there is an upward pressure on the currency due to excess demand, the central bank can keep on selling the domestic currency and buying foreign exchange, thus maintaining the exchange rate.

Increases in interest rates

The central bank can increase interest rates, which attract financial investments from other countries (see Section 16.1). This leads to a higher demand for the domestic currency, shifting the demand for boples back to D_1 in Figure 16.4(a). However, increases in interest rates involve contractionary monetary policy and may lead to a recession in the domestic economy.

Borrowing from abroad

If the country borrows from abroad, its loans will come in the form of foreign exchange, which when converted into boples will cause an increase in the demand for boples and hence a rightward shift in the demand curve toward D_1. However, extensive borrowing from abroad comes with a number of costs (see Chapters 19 and 20).

Efforts to limit imports

The government could use policies to limit imports, because this reduces the supply of the domestic currency (since it reduces demand for foreign exchange needed to buy imports), causing a leftward shift in the currency supply curve. In Figure 16.4(b), this appears as a shift from S_1 to S_2, with the exchange rate remaining fixed at 2.00 dollars = 1 bople (point B, given by the intersection of D_2 with S_2). To limit imports, governments can use (a) contractionary fiscal and monetary policies, which lower aggregate demand, lower incomes, and therefore result in fewer imports; (b) trade protection trade policies, which work to directly lower the quantity of imports that can enter the country. However, contractionary policies may lead to recession, while trade protection comes with numerous disadvantages, including the possibility of retaliation by trading partners, which would result in lower exports (see Chapter 14).

Exchange rates are *always* determined by demand and supply of a currency in both floating and fixed exchange rate systems. Whereas in a floating exchange rate system the 'price' of a currency adjusts freely to changes in supply and demand, in a fixed exchange rate system the currency supply and demand are forced to adjust to the predetermined 'price', or fixed exchange rate, through central bank and government intervention that manipulates them.

Changing the fixed exchange rate: devaluation and revaluation

If a country experiences serious difficulties in maintaining the fixed exchange rate, a different fixed rate can be set. If the currency value is higher than what can be maintained through intervention, the government may change it to a new, lower value; this is called **devaluation** of the currency. If, on the other hand, the currency has a lower value than can be maintained by intervention, the government may set a new higher value; this is called **revaluation**.

As an example of devaluation suppose 2 US dollars exchange for 1 British pound; the dollar devalues and, at the new fixed rate, 3 dollars exchange for 1 pound. Before the devaluation, 2 dollars were needed to buy 1 pound; now 3 dollars are needed, because the dollar has lost some of its value. This is equivalent to a revaluation of the pound relative to the dollar (the pound increases in value).

Like depreciation, devaluation results in cheaper exports to foreigners and more expensive imports for domestic residents, giving rise to more exports and fewer imports. Like appreciation, revaluation leads to more expensive exports to foreigners and cheaper imports for domestic residents, and therefore fewer exports and more imports.

Historically, a system of fixed exchange rates was in place until 1973. In the period 1879–1934, the fixed rate system was known as the 'gold standard', as countries fixed their exchange rates relative to the value of gold. In the period 1944–1973, the fixed rate system came to be known as the Bretton Woods system, which no longer tied currencies to gold, and permitted periodic devaluations or revaluations. Under this system, when any country revalued/devalued its currency, it did so not against just one other currency but against all other currencies simultaneously, since all currencies were fixed against each other. This is analogous to the appreciation and depreciation that takes place under flexible exchange rates.

Today, the closest to fixed exchange rates are pegged exchange rates, involving a currency that is fixed against one other currency. While this is similar to a fixed exchange rate, it is considered by the International Monetary fund (IMF) to be a type of managed system, so will be considered below.[2]

TEST YOUR UNDERSTANDING 16.5

1 Describe a fixed exchange rate system.

2 a Using a diagram, explain some methods that central banks and governments can use to maintain a fixed exchange rate.

  b Why do each of these lead to problems?

3 a Distinguish between devaluation and revaluation.

  b When is each of these undertaken by a government?

4 Distinguish between

  a depreciation and devaluation, and

  b appreciation and revaluation.

5 Outline why it is easier to maintain a fixed exchange rate when there are upward pressures on the value of a currency than when there are downward pressures.

Managed exchange rates

Understanding managed exchange rates

In between the two extremes of floating exchange rates and fixed exchange rates is the system of **managed exchange rates**, also known as the *managed float*. Combining elements of both, though closer to floating exchange rates, this is the current system, in use since 1973. Exchange rates are for the most part free to float to their market levels over long periods of time; however, central banks periodically intervene to stabilise them over the short term.

The objective of central bank intervention is to prevent large and abrupt fluctuations in exchange rates that could arise if currencies were left entirely to market forces. Large and abrupt exchange rate changes disrupt the orderly flow of international trade and create uncertainties that undermine investment and economic activity. Under managed exchange rates, the currency is supposed to move towards its long-term equilibrium position determined by the market. Central banks intervene so that this adjustment can occur in a smooth and orderly way, without major and abrupt fluctuations that may destabilise the economy.

Intervention mainly takes the form of buying and selling of currencies by the central bank, influencing currency demand and supply. In addition, central banks may change interest rates, which also impact upon exchange rates. Very infrequently (mainly in the event of a severe disequilibrium where there is a strong downward pressure on the value of a currency) governments may have to resort to contractionary macroeconomic policies or protectionist measures (as in the case of fixed exchange rates).

Pegging exchange rates

A number of countries peg (i.e. fix) their currencies to the US dollar, and float together with it, while a few other economies peg their currencies to the euro. The pegged currency is allowed to fluctuate only within a narrow range above and below a target exchange rate relative to the dollar or the euro, so that if the actual exchange rate hits the upper or lower limit of the range, the central bank intervenes to keep it within the limits.

[2] The IMF, discussed in Chapter 20, overseas the global financial system. See https://www.imf.org/external/pubs/ft/wp/2009/wp09211.pdf

To see how this works, suppose Bopland decides to peg the bople to the US dollar, as shown in Figure 16.5. The target exchange rate chosen is 2 dollars = 1 bople; the bople is allowed to fluctuate up to a maximum of 2.10 dollars = 1 bople, and a minimum of 1.90 dollars = 1 bople. Suppose that market forces cause the exchange rate to drop to 1.90 dollars = 1 bople, due to a fall in the demand for boples from D_1 toward D_2. At that point the Bank of Bopland (the central bank) will intervene by buying boples (and selling dollars), so that the demand-for-boples curve will stop shifting leftward and the bople stops falling. If the bople increases in value and hits the maximum of 2.10 dollars = 1 bople because of an increase in the demand for boples from D_1 toward D_3, the central bank will intervene by selling boples (and buying dollars) to prevent a further rise in the value of the bople.

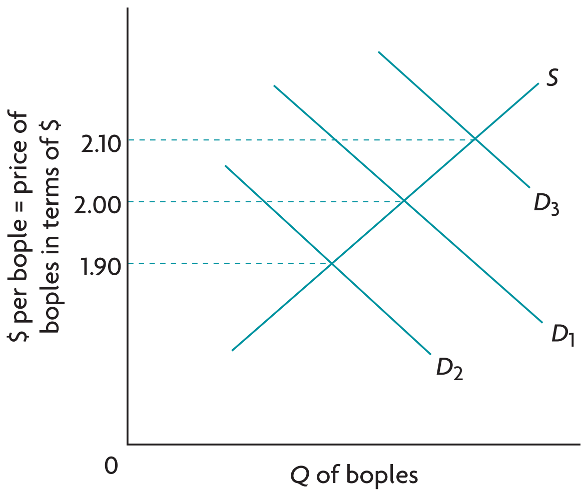


Figure 16.5: Illustrating a pegged currency

A pegged currency combines fixed and floating exchange rates, because the pegged currencies are fixed within the specified range of the US dollar (or the euro), and they float in relation to all other currencies, together with the dollar (or the euro). Pegging a currency stabilises its exchange rate in relation to the currency to which it is pegged, preventing abrupt or strong fluctuations. Countries that peg their currencies to the US dollar experience exchange rate stability relative to the dollar as well as relative to each other, and this facilitates trade flows with the United States as well as between the countries with pegged currencies. In addition, some economies that are strong financial centers (such as Hong Kong; see Real world focus 16.2) peg their currency to the US dollar.

Under the managed float, exchange rates are determined mainly through market forces, but with periodic intervention by central banks aiming to smooth out abrupt fluctuations. Intervention takes mainly the form of the buying and selling of official reserves. Some economies peg their currencies to the dollar or euro; pegged currencies are fixed in relation to the dollar or euro, and float in relation to all other currencies.

Consequences of overvalued and undervalued currencies

An **overvalued currency** has a value that is too high relative to its equilibrium free market value. Its exchange rate has been set at a higher level than the equilibrium market exchange rate. An **undervalued currency** has a value too low relative to its equilibrium free market value; its exchange rate is low relative to the one the market would have determined.

Overvalued and undervalued currencies cannot come about in a freely floating exchange rate system, where exchange rates are determined purely by free market forces. However, they can and do often occur in fixed and managed exchange rate systems.

We can use Figure 16.1(a) to illustrate both overvalued and undervalued currencies. The market determines a price of US dollars at 0.67 euro = 1 dollar. If the US central bank (known as the Federal Reserve) wanted to overvalue the US dollar, it could try to maintain a price in terms of the euro above the equilibrium, such as at 0.80 euro = 1 dollar. In the case of an undervaluation, the central bank would select a price of the dollar in terms of the euro below the equilibrium price, such as at 0.50 euro = 1 dollar. The overvaluation or undervaluation of the currency can be achieved by central bank and government interventions that maintain the exchange rate at the selected level (or range of levels, as when a currency is pegged).

There are a number of advantages that may arise from overvaluation and undervaluation of currencies, though these generally come with costs.

Overvalued currencies

Most developing countries have at one time or other had overvalued exchange rates. If an exchange rate is overvalued, imports become cheaper. The main reason for overvaluing their exchange rates is that many developing countries have wanted cheap imports of

REAL WORLD FOCUS 16.2

Hong Kong dollar pegged to the US dollar

Hong Kong has been pegging its currency to the US dollar since 1983. The Hong Kong dollar is pegged at HK$7.80 to US$1and is allowed to fluctuate within a band of ten cents relative to the US dollar (HK$7.75 to HK$7.85). The Hong Kong Monetary Authority (HKMA, or central bank) changes interest rates mainly in order to support the peg within this narrow band. For example, interest rates rise when the Hong Kong dollar falls to the lower range of the band, and fall when the Hong Kong dollar rise to the higher edge of the band. Apart from this, interest rates in Hong Kong can change only as long as the Hong Kong dollar remains within the predefined narrow band relative to the US dollar. In effect the HKMA is unable to carry out an independent monetary policy.

Figure 16.6: 100 Hong Kong dollar with 100 US dollar

The HKMA also has another weapon to defend the Hong Kong dollar in case this begins to weaken below the lower edge of the band. It holds US$440 billion in reserve assets (foreign exchange reserves), which amounts to 1.4 times its GDP. This is the largest amount of *per capita* foreign exchange reserves in the world.

Sources: *William Pesek, 'Time to scrap Hong Kong's currency peg', Nikkei Asian Review, 28 May 2018, Stefan Gerlach, 'Hong Kong has been well served by pegged exchange rate for 35 years', South China Morning Post, 8 May 2018*

Applying your skills

1. Use an exchange rate diagram to explain how the HKMA is able to keep the Hong Kong dollar pegged to the US dollar within the narrow band of ten cents by use of
   - a interest rates, and
   - b foreign exchange reserves.
2. Using your knowledge about fixed exchange rates, discuss some advantages and disadvantages of the HKMA pegged exchange rates.

capital goods, raw materials and other inputs for use in manufacturing industries, to speed up industrialisation.

However, overvalued exchange rates come with many disadvantages. One of the most important of these is that exports become more expensive, thus hurting domestic exporters. Increased imports and reduced exports lead to a worsening current account balance (referred to above, and explained below, Section 16.4), resulting in payments difficulties. In addition, by increasing imports, overvalued exchange rates can hurt domestic producers who have to compete with artificially low-price imports, with negative consequences for domestic employment and resource allocation. Overvaluation has often resulted in the need for countries to devalue or depreciate their currencies to correct the overvaluation.

Undervalued currencies

When a currency is undervalued, exports become less expensive to foreign buyers, while imports become more expensive domestically. Some developing countries have used undervaluation as a method to expand their export industries, expand their economies and therefore also increase their employment levels. Achieving these objectives by means of an undervalued currency is considered to involve the creation of an unfair competitive advantage compared to other countries that do not undervalue their currencies, and which suffer the consequences of increased imports and lower exports. Currency undervaluation is therefore considered to be a kind of 'cheating'. In the context of a managed float, undervalued currencies are sometimes referred to as a 'dirty float'. A disadvantage of an overvalued currency is that it can lead to cost-push inflation due to the higher price of imports. Correction of the undervalued currency would involve revaluation or appreciation of the currency.

TEST YOUR UNDERSTANDING 16.6

1. a In what ways is the managed float an exchange rate system that lies between fixed and floating exchange rate systems?

   b Outline why is it closer to floating exchange rates.

2. Describe the reasons for government intervention in a managed float.

3. Outline the meaning of pegging a currency in the context of the managed float system.

4. a Distinguish between overvalued and undervalued exchange rates.

   b State the reasons for overvaluing or undervaluing a currency.

   c Outline the disadvantages of each.

5. a Explain why overvalued and undervalued exchange rates do not arise in a freely floating exchange rates system.

   b Outline why undervalued currencies are sometimes referred to as a 'dirty float'.

6. Research and find one or more countries that peg their currency to another currency. Based on your findings, determine the reasons for the peg. Describe the advantages and disadvantages of the pegged currency in your example.

Comparing and contrasting fixed and floating exchange rate systems (HL only)

This topic will be discussed in Chapter 17, Section 17.2, after you have learned about the balance of payments.

16.4 The balance of payments

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the balance of payments in terms of the current account, the capital account and the financial account, and the items that compose each of the three accounts (AO2)
- explain the interdependence between the three accounts in terms of a zero balance, the balance between credits and debits, and the balance between deficits and surpluses (AO2)
- calculate items of the balance of payments given a set of data (AO4)

The role of the balance of payments

Innumerable transactions between residents of different countries involve activities like importing and exporting, travel, investments in stocks and bonds, investments by multinational corporations, buying property, sending or receiving gifts, and in general, anything that gives rise to a flow of money across international boundaries. All these transactions are recorded in the balance of payments:

The **balance of payments** of a country is a record (usually for a year) of all transactions between the residents of the country and the residents of all other countries. Its role is to show all payments received from other countries, called **credits**, and all payments made to other countries, called **debits**. In the course of a year, all inflows of payments (credits) must exactly equal the outflows of payments (debits); the sum of all credits is equal to the sum of all debits.

Why do countries record all their foreign transactions; why do they keep records of all possible inflows and outflows of money? The reason is that when money flows into or out of a country, it involves the exchange of different national currencies. These currency exchanges are an important part of understanding the balance of payments.

Relating the demand and supply of a currency to the balance of payments

We will consider the balance of payments accounts of Bopland, whose national currency is the bople. All the inflows of money from abroad into Bopland, or all credits, can only be made if foreigners buy boples; therefore *credits represent a foreign demand for boples*, corresponding to a foreign supply of all other currencies given up to buy boples. Outflows of money from Bopland to other countries, or *debits, represent a Boplander supply of boples*, corresponding to Boplander demand for foreign currencies.

In the balance of payments accounts of a country, all credits (inflows of money into the country) create a foreign demand for the country's currency; and all debits (outflows of money from the country) create a supply of the domestic currency.

The structure of the balance of payments

Table 16.3 shows Bopland's balance of payments for the year 2019. The balance of payments consists of three accounts: the current account, the capital account and the financial account. Each of the items in the three accounts is accompanied by a plus or minus sign: the plus sign denotes *credits* (money inflows) and the minus sign denotes *debits* (money outflows).

For each of these three accounts, when the credits and debits are added up, the sum is a positive or negative number, depending on whether the credits are greater than the debits (a positive sum) or the debits are greater than the credits (a negative sum).

A surplus in an account occurs whenever a balance has a positive value, meaning that credits are larger than debits (there is an excess of credits).

A deficit in an account occurs whenever a balance has a negative value, meaning that debits are larger than credits (there is an excess of debits).

Current account	
1 Exports of goods	+40
2 Imports of goods	−65
Balance of trade in goods (items 1+2)	**−25**
3 Exports of services	+25
4 Imports of services	−15
Balance of trade in services (items 3+4)	**+10**
Balance of trade in goods and services (items 1+2+3+4)	**−15**
5 Income (inflows minus outflows)	−6
6 Current transfers (secondary income) (inflows minus outflows)	+1
Balance on current account (items 1+2+3+4+5+6)	**−20**
Capital account	
7 Capital transfers (inflows minus outflows)	+0.7
8 Transactions in-produced, non-financial assets (inflows minus outflows)	+0.3
Balance on capital account (items 7+8)	**+1**
Financial account	
9 Foreign direct investment (FDI; inflows minus outflows)	+23
10 Portfolio investment (inflows minus outflows)	−4
11 Reserve assets (official reserves)	+1
12 Official borrowing	−1
Balance on financial account (items 9+10+11+12)	**+19**
Balance (sum of all items from 1 to 12)	**0**

Table 16.3: Balance of payments of Bopland, 2019 (in billions of boples)

The current account

Balance of trade in goods

Items 1 and 2 show the value of exports and the value of imports of goods respectively. Exports are the sale of goods to other countries, for which payment is received in boples; therefore exports are a credit and have a plus sign. Imports are the purchase of goods from other countries, for which payment is made in foreign currencies (generating a supply of boples); imports are therefore a debit and have a minus sign. The **balance of trade in goods** is calculated by subtracting imports from exports, or what is the same thing, adding items 1 and 2 (note that we must take into account the minus sign), so the balance of trade is $40 - 65 = -25$ billion boples, indicating a negative balance of trade in goods, or a *deficit* in this balance.

Balance of trade in services

Items 3 and 4 are analogous to items 1 and 2, only they involve the value of exports and value of imports of services. Services include a variety of activities, such as insurance, tourism, transportation and consulting. When foreigners visit Bopland as tourists, Bopland is exporting tourism services; similarly, when foreigners buy insurance from Boplander companies, this represents exports of insurance services. When Boplanders visit other countries as tourists, or buy insurance from other countries, they are importing tourism and insurance services. Table 16.3 shows that Bopland's exports of services are larger than its imports of services, and so the **balance of trade in services** (items 3 + 4) is +10 billion boples. Bopland therefore has a *surplus* in its balance of trade in services.

Balance of trade in goods and services

The next line states 'Balance of trade in goods and services', which is the sum of all the items 1-4. This is often referred to as the 'trade balance' or 'balance of trade' for short, and includes the value of all exports minus the value of all imports. You may note that this corresponds to what we call the 'net exports' component of GDP, abbreviated as $X - M$. Bopland has a negative trade balance (–15 billion boples).

Income

Item 5, or **income** refers to all inflows into Bopland of rents, interest and profits from abroad, minus all outflows of rents, interest and profits. Boplanders may earn income abroad, if they own rental property abroad that earns rental income, or have bank accounts abroad that earn interest, or if they own stocks in another country that earn dividend income, or if they own a subsidiary of a multinational corporation that earns profits. Whatever income flows into Bopland from abroad is a credit, while whatever income flows out of Bopland is a debit. In Bopland, income outflows are greater than income inflows, leading to a value of -6 billion boples for income.

Current transfers

Item 6, or **current transfers** refers to inflows into Bopland due to transfers from abroad like gifts, remittances (money sent home to relatives by Boplanders living abroad), foreign aid, and pensions, minus outflows of such transfers to other countries. This item is positive (+1 billion boples), indicating that credits are greater than debits, as Bopland receives more transfers from abroad than it makes to other countries.

Balance on current account

When we add up all the items in the current account (items 1–6), we get the *balance on current account* known also as the **current account balance**. Bopland has a **current account deficit** of -20 billion boples; the debits (outflows) are larger than the credits (inflows) by this amount. (If the credits were larger than the debits Bopland would have had a **current account surplus**.)

From the point of view of demand and supply of boples, *the deficit on this account means there is an excess supply of the currency in the foreign exchange market*: the quantity of boples supplied (debits created by Boplanders) is larger than the quantity of boples demanded (credits created by foreigners). (The quantity of boples demanded to make the credits possible is equal to the sum of all the credits, or $+40 + 25 + 1 = +66$ billion boples; whereas the quantity of boples supplied to make the debits possible is equal to the sum of all the debits, or $-65 - 15 - 6 = -86$. Adding the credits to the debits we have $+66 - 86 = -20$ billion boples, which is *excess supply of boples*, or the deficit on the current account.)

In general, the trade balance (meaning the balance of trade in goods and services) is the most important part of the current account in most countries. Therefore a deficit on the current account is usually due to an excess of imports of goods over exports, whereas a surplus on the current account is usually due to an excess of exports of goods over imports.

The current account of the balance of payments is the sum of: (i) the balance of trade in goods; (ii) the balance of trade in services; (iii) income inflows minus outflows; and (iv) current transfer inflows minus outflows. The most important part of the current account in most countries is the balance of trade in goods and services (i + ii).

The capital account

The *capital account* consists of two items. In item 7, **capital transfers**, include inflows minus outflows for such things as debt forgiveness (when debt is cancelled), non-life insurance claims, and investment grants (money given as a gift by governments to finance physical capital).

Item 8, transactions in **non-produced, non-financial assets**, consist mainly of the purchase or use of natural resources that have not been produced (land, mineral rights, forestry rights, water, fishing rights, airspace and electromagnetic spectrum). It includes all inflows of funds into Bopland (credits) minus outflows of funds from Bopland (debits) due to such transactions.

The sum of items 7 and 8 give the *balance on capital account*, in which Bopland has a surplus of +1 billion boples, meaning that the credits (payment inflows) are more than the debits (payment outflows). The quantity of boples demanded is larger than the quantity of boples supplied. Therefore *a surplus on an account indicates there is an excess demand of the currency in the foreign exchange market*.

In general, the capital account is relatively unimportant in terms of size compared to the other two accounts

The capital account of the balance of payments of a country is composed of inflows minus outflows of funds for capital transfers and transactions in non-produced, non-financial assets. The capital account is relatively small compared to the current account and financial account.

The financial account

The financial account consists of four items.

Foreign direct investment (FDI)

Item 9 deals with **foreign direct investment** (to be studied in Chapter 20). This includes investments in productive facilities, consisting of physical capital, such as buildings and factories, undertaken by multinational corporations. The figure for this item includes inflows due to direct investment by foreigners in Bopland (credits) minus outflows due to Boplander investment abroad (debits). Bopland accepted more direct investments by foreigners than it made in other countries, by the amount of 23 billion boples.

Portfolio investment

Item 10, **portfolio investment**, shows investments in financial capital (such as stocks and bonds). In Bopland, inflows (credits) for the purchase of stocks and bonds were less than outflows (debits) for the same purpose (–4 billion boples).

(You may note the distinction between inflows or outflows of funds due to the purchase of assets, and inflows or outflows of funds due to income generated by the purchase of assets. If a multinational corporation decides to invest in Bopland by purchasing physical capital, there results an inflow of funds into Bopland appearing as a credit in Bopland's *financial* account. If the owners of the multinational corporation decide to take their profits out of Bopland and back to the home country, there is an outflow of funds from Bopland appearing as a debit in Bopland's *current* account.)

Reserve assets

Item 11, **reserve assets** (or official reserves), refers to foreign currency reserves that the central bank can buy or sell to influence the value of the country's currency . Suppose the Central Bank of Bopland holds reserves of US dollars. If it sells dollars, it does so by buying boples. This is an inflow of boples, appearing as a credit in the financial account. Table 16.3 shows the Central Bank to have bought 1 billion boples, appearing with a plus sign (a credit). (If the central bank had sold boples by buying dollars, this would be an outflow of boples and would appear as a debit in the financial account.)

Official borrowing

Official borrowing refers to government borrowing from abroad. Inflows of funds into Bopland due to borrowing by the Boplander government from foreign lenders (foreign government debt, to be discussed in Chapter 19; also Chapter 11 at HL) appear as credits. Similarly, Boplander loans to foreign governments lead to an outflow of funds from Bopland appearing as debits. The figure of –1 billion boples indicates that Bopland is a net lender (outflows are greater than inflows).

Balance on financial account

The *balance on financial account*, given by the sum of items 9, 10, 11 and 12 shows a surplus of 19 billion boples (the credits are more than the debits). Therefore there is an excess demand of 19 billion boples in the foreign exchange market for this account.

> The **financial account** of the balance of payments consists of inflows minus outflows of funds for (i) foreign direct investment; (ii) portfolio investment; (iii) reserve assets and (iv) official borrowing.

Two clarifications

Errors and omissions

In the real world, it is extremely difficult (if at all possible) to record every single transaction between a country and all other countries, and some of these go unrecorded. However, since the sum of all credits must equal the sum of all debits, it is necessary for actual accounts to include an item creating this equality. Therefore, the real-world balance of payment accounts include an item called 'errors and omissions' to create an equality between the sum of credits and the sum of debits. It is simply a statistical 'trick' that does not affect our analysis of the balance of payments, and so for simplicity is not included here.

The capital account and the financial account

Economists often use the term 'capital account' to refer to both the capital and financial accounts that appear in Table 16.3. However, since 1997, countries around the world are increasingly using the classification system shown in Table 16.3 in their balance of payments. Therefore if you come across the expression 'capital account' in your general reading, you should be aware that reference is being made to what is actually the 'financial account', together with the relatively unimportant 'capital account'.

TEST YOUR UNDERSTANDING 16.7

1 Define the balance of payments. Describe its role.

2 a Define debits and credits, relating these to the demand for or supply of a currency in the foreign exchange market.

 b Provide some examples of money flows into and out of the country where you live, and explain whether these are credits, or debits.

3 Explain, providing examples,

 a the four components of the current account,

 b the two components of the capital account, and

 c the four components of the financial account.

4 a Outline the meaning of a 'deficit' and a 'surplus' on an account. Explain in terms of debits and credits.

 b Explain the meaning of a deficit or surplus in the current account. Identify what items in the balance of payments are most likely responsible for a deficit or surplus.

Calculating elements of the balance of payments

Given the components of the balance of payments and their relationships, we can calculate various elements in the balance of payments. You are given some exercises in Test your understanding 16.8.

TEST YOUR UNDERSTANDING 16.8

Answer the questions below based on the table showing the balance of payments accounts of Lakeland (2019, billion Lkl).

Current account	
Exports of goods	+ 310
Imports of goods	– 525
Balance of trade in goods	
Exports of services	+ 52
Imports of services	– 71
Balance of trade in services	
Income	+ 25
Current transfers	+ 73
Balance on current account	
Capital account	
Capital transfers	– 3
Transactions in non-produced, non-financial assets	+ 7
Balance on capital account	
Financial account	
Direct investment	+ 107
Portfolio investment	+ 29
Reserve assets	– 7
Official-borrowing	
Balance on financial account	
Balance	

1. Fill in the blanks in the table. Check your results by using the relation: current account = capital account + financial account.
2. Identify which of the three accounts are in surplus and which in deficit. Explain your answer using the concepts of debits and credits.
3. Is Lakeland experiencing a balance of payments deficit or a balance of payments surplus?
4. Outline whether or not it is possible that the value of the Lkl is determined in a freely floating exchange rate system? Explain.

The interdependence between the accounts

The meaning of 'zero balance' in the balance of payments

In the balance of payments, the sum of all the items is always zero. This is another way of saying that the *sum of all credits always balances with the sum of all debits*. In addition it is another way of saying that *deficits are matched by surpluses*.

As you can see in Bopland's case, the deficit in the current account of −20 billion boples is exactly matched by the surplus in the combined capital and financial accounts: −20 billion boples ⇔ + 1 + 19 billion boples. In other words, the excess supply of boples in the current account, which is in deficit, is exactly matched by an excess demand for boples in the remaining two accounts, which altogether are in surplus.

We turn now to an explanation of this point.

Why is there a zero balance in the balance of payments?

The question arises, why does the balance of payments have a zero balance?

There is a very simple answer to this question. You may recall from our discussion above that all credits in the balance of payments create demand for a currency, and all debits create supply of the currency. You may also recall that *exchange rates are always determined by currency demand and supply*, in both floating and fixed exchange rate systems. The same is true for managed exchange rates since these are floating exchange rates with some intervention to influence currency demand or supply. It follows then that at the point where an equilibrium exchange rate is determined, in other words *at the point where currency demand equals currency supply, it is also the case that the sum of credits is equal to the sum of debits. Therefore deficits must match surpluses*.

How does this work *in practice*? We turn to this question next.

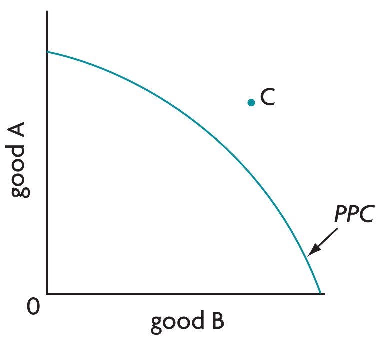
Why the current account and financial account are interdependent

When a country trades with other countries, its imports of goods and services are unlikely to be equal to its exports. If imports are greater than exports, it has a deficit in its trade balance, and since this is the most important component of the current account, it also likely has a current account deficit. Using the production possibilities model, we can see this in Figure 16.7(a). The country's *PPC* defines the maximum it can produce, but the country is consuming at a point outside its *PPC*, such as point C, because it is importing more than it is exporting.

(HL students note that achieving a point outside the *PPC* by means of a trade deficit is very different from achieving such a point by specialisation and trade according to comparative (or absolute) advantage (Chapter 14). The theory of comparative advantage presupposes that imports and exports balance each other, so there is no trade deficit or trade surplus: a point outside the *PPC* is achieved because of an increase in allocative efficiency.)

If there is a current account deficit, there must be a financial account surplus (the capital account being very small), which provides it with the foreign exchange it needs to pay for the excess of imports over exports. The surplus on the financial account may arise from investments in physical or financial capital by foreigners including loans from foreigners. It follows, then, that a deficit in the current account is matched by a surplus in the financial account (along with the unimportant capital account).

a With a trade deficit, country consumes outside its *PPC*



b With a trade surplus, country consumes inside its *PPC*

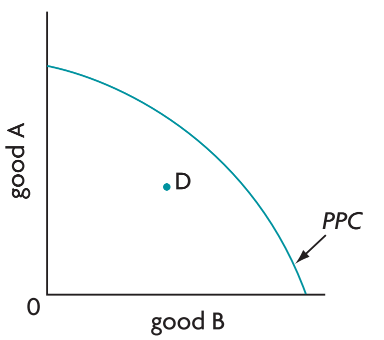


Figure 16.7: Using a *PPC* to illustrate a trade deficit and a trade surplus

If the economy's exports of goods and services are greater than its imports, it has a surplus in its current account, meaning it is buying from foreigners less than what it sells to them. While it is producing somewhere on the *PPC*, it is consuming less, so the output available for domestic consumption is at a point inside its *PPC*, such as D in Figure 16.7(b). The difference between what it consumes and what it produces is the excess of exports over imports.

When there is a surplus on the current account, the country is accumulating foreign exchange (as it earns more foreign exchange from exports than it pays out to buy imports), which it can use to buy assets abroad (direct or portfolio investments, including loans to other countries). It follows, then, that a surplus in the current account is matched by a deficit in the financial account.

In fact, most economists believe that the surplus or deficit of a financial account of a country very often *is the result* of what is happening in the current account. If there is a deficit in the current account, the financial account is a reflection of the need to finance that deficit; if there is a surplus in the current account, the financial account reflects investments in foreign countries undertaken to dispose of the extra foreign exchange.

A current account deficit means a country consumes more than it produces; and it pays for extra output consumed through a financial account surplus. A current account surplus means a country consumes less than it produces, and part of the income generated from the sale of extra output produced corresponds to a financial account deficit.

The meaning of 'imbalance' in the balance of payments

Since the balance of payments must always balance, why do we often hear the expressions 'balance of payments deficit' or 'balance of payments surplus'? A balance of payments deficit means there is a deficit in the combined current, capital and financial accounts (plus errors and omissions), *excluding central bank intervention*. A balance of payments surplus means there is a surplus in the combined three accounts (plus errors and omissions), *excluding central bank intervention*. Bopland, for example, has a balance of payments deficit:

current account deficit (− 20 billion) + capital account surplus (+ 1 billion) + financial account surplus excluding the central bank purchase of 1 billion (+ 18 billion) = −1 billion = balance of payments deficit

The reason for central bank intervention can be found in just this deficit. By buying up excess boples of 1 billion, the central bank creates a balance in the balance of payments.

Such imbalances (deficits or surpluses) occur in virtually all countries all the time, but a 'balance' is always created in the sense that debits are made to equal credits. There are many ways to do this, which involve either reliance on market forces or on government intervention. HL students will see how this is accomplished in Chapter 17.

TEST YOUR UNDERSTANDING 16.9

1 Explain why
   a a trade surplus means that a country is consuming less than it is producing
   b a trade deficit means that a country is consuming more than it is producing.
2 Explain the relationship between the current account, capital account and financial account.
3 Using the balance of payments accounts for Lakeland appearing in Test your understanding 16.8, draw a PPC diagram to show how much Lakeland is consuming in relation to how much it is producing.
4 Explain why current account deficits are likely to be roughly matched by financial account surpluses, and current account surpluses to be roughly matched by financial account deficits.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Select a country you are interested in and examine how the exchange rate of its currency has developed over the course of the last several years, noting when it has been appreciating and when it has been depreciating. Try to identify the most important factors that have led to exchange rate changes. Have these factors led to changes in currency demand, or currency supply or both?
2 Select a country you are interested in that has a managed exchange rate. Try to identify the objectives of the government or central bank intervention in the foreign exchange market to manage the exchange rate; is intervention intended to lead to currency appreciation or depreciation relative to the free market value? Why does the government or central bank want the currency to appreciate or depreciate? What methods does the government or central bank use to achieve the objectives regarding the exchange rate? Can you identify any important consequences of exchange rate management in the domestic economy?
3 Choose one or more countries that are major trading partners of the country you live in. Does your country of residence have a current account surplus or deficit with this country or countries? To what extent is the current account surplus or deficit due to a trade surplus or deficit? Are there are other components of the current account, such as income (possibly profit income or remittances), that play an important role in determining the current account balance?

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 17

Further topics on exchange rates and the balance of payments (HL only)

BEFORE YOU START

- As you have learned, a current account deficit/surplus is driven by a country's balance of trade. What do you think could be the advantages or disadvantages of running a current account deficit or surplus? What if the deficit or surplus is persistent over many years?
- A monetary union means participating countries use a common currency while sharing a common central bank. What might be some advantages and disadvantages of this level of economic integration?

This chapter will explore more closely some of the topics on international monetary flows that were introduced in Chapter 16. We will examine how surpluses and deficits in the current and financial accounts are likely to impact upon exchange rates. We will evaluate fixed and floating exchange rate systems, as well as monetary union, which in some ways is like a system of permanently fixed exchange rates. Finally, we will study the consequences of imbalances in the balance of payments, as well as how countries deal with the potentially serious issue of persistent current account deficits.

Currencies of different countries

17.1 How the current account and the financial account are related to exchange rates

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the relationship between the exchange rate and the current account (AO2)
- draw a diagram showing the relationship between the current account balance and the exchange rate (AO4)
- explain the relationship between the exchange rate and the financial account (AO2)

The current account and exchange rates

The current account and exchange rates in a floating exchange rate system

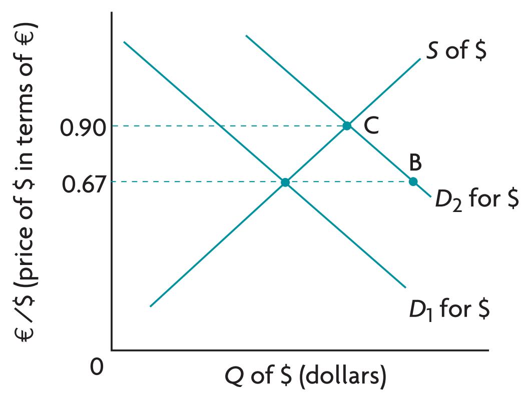
As we know, under floating exchange rates, there is no government or central bank intervention, and the market determines the equilibrium exchange rate. (In Table 16.3, there would be no buying or selling of currencies by the central bank, and the entry for item 11 on reserve assets would be zero.)

In Figure 17.1(a), at the initial equilibrium exchange rate of 0.67 euro = 1 dollar, the quantity of dollars demanded is equal to the quantity of dollars supplied. In equilibrium, the sum of credits is equal to the sum of debits in the US balance of payments. Similarly, in Figure 17.1(b), at the (identical) exchange rate of 1.5 dollars = 1 euro, the quantity of euros demanded is equal to the quantity of euros supplied, so the sum of credits is equal to the sum of debits for the euro zone countries as well.

If the euro zone countries' demand for imports from the United States increases, in Figure 17.1(a) the demand for dollars increases and the demand curve shifts to the right from D_1 to D_2. At the initial equilibrium exchange rate of rate 0.67 euro = 1 dollar, there is an excess demand for dollars equal to the horizontal distance between A and B. The United States now has a *surplus on its current account*, and since nothing has changed in the remaining accounts, it has *excess credits in its balance of payments*. An imbalance has therefore been created.

Part (b) shows that the increased demand for imports from the United States has caused the supply of euros to increase, and the supply curve shifts to the right from S_1 to S_2. At the initial equilibrium exchange rate of 1.5 dollars = 1 euro, there is an excess supply of euros equal to the horizontal distance between D and E. The euro zone countries have *a current account deficit*, corresponding to *excess debits in their balance of payments*. The current account surplus in the United States corresponds to a current account deficit in the euro zone countries.

a Current account surplus causes appreciation



b Current account deficit causes depreciation

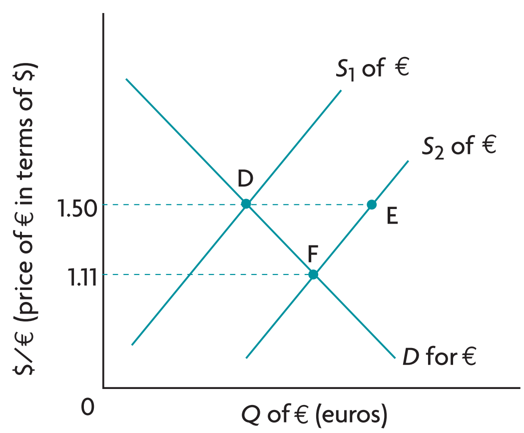


Figure 17.1: Current account and exchange rates

We know that under freely floating exchange rates, market forces cause the exchange rate to change: the dollar appreciates to 0.90 euro = 1 dollar, which is equivalent to the depreciation of the euro to 1.11 euro = 1 dollar. In the United States the dollar appreciation causes imports to increase and exports to decrease until the current account surplus is eliminated. In the euro zone countries the euro depreciation causes imports to fall and exports to increase until the current account deficit is eliminated.[1] This leads to an important conclusion:

> Under floating exchange rates, when there is a deficit in the current account, market forces create a downward pressure on the currency exchange rate. When there is a surplus in the current account, market forces create an upward pressure on the currency exchange rate. As a result, exchange rate changes automatically eliminate current account deficits and surpluses, and create a balance in the balance of payments.

The current account and exchange rates in a managed exchange rate system

The managed float is close to a freely floating exchange rate only it includes periodic interventions by the central bank to influence exchange changes (see Chapter 16). The most common intervention involves buying and selling of reserve currencies (reserve assets), as with the Central Bank of Bopland's intervention. As we saw in Table 16.3 Bopland has a deficit in its balance of payments, meaning there are excess debits due to an excess supply of boples. In a freely floating system, the central bank would not have intervened, and *the bople would have depreciated*. However, the Central Bank of Bopland sold dollars and bought 1 billion boples, creating a credit of that amount, therefore offsetting the deficit of −1 billion boples. The result was to *avoid depreciation and maintain the bople's exchange rate*. An alternative would have been for the central bank to create a smaller credit (say 0.5 billion boples), in which case the bople would have depreciated, but less than under a freely floating system.

The current account and exchange rates in a fixed exchange rate system

Suppose now that Bopland has a fixed exchange rate. If excess debits persist year after year, the Central Bank of Bopland can keep selling dollars and buying boples, creating the necessary credits to match the excess debits, thus maintaining the fixed exchange rate. But at some point, the Bank of Bopland will run out of dollars to sell. The government or central bank must therefore find ways to increase credits or decrease debits to maintain a balance. To increase credits, the central bank can increase interest rates, thus attracting foreign financial investments; or the government can borrow from abroad; both actions increase credits in the financial account. The government could limit imports (through contractionary fiscal and monetary policies or trade protection), or it could impose exchange controls (limits to the amount of domestic currency that can be exchanged for foreign currencies); both these actions decrease the debits.

These measures are exactly the same as those needed to maintain the fixed exchange rate, studied in Chapter 16, Section 16.3. It follows that

> In a fixed exchange rate system, the balance of payments are made to balance by policies that change currency demand or supply in order to keep the exchange rate fixed.

You can see, then, that the idea of a zero balance in the balance of payments is very closely connected to exchange rates. The reason is that *everything that is recorded in the balance of payments creates a demand for or supply of a domestic currency*. Since the value of the domestic currency is determined by currency supply and demand (in all exchange rate systems), it follows

[1] To see how this happens, consider the following chain of events. Given the initial change in the demand for dollars, the appreciation of the dollar makes US exports more expensive and imports from euro zone countries cheaper; while the depreciation of the euro makes euro zone exports cheaper and imports from the US more expensive. More expensive US exports are in effect more expensive imports from the United States for euro zone countries, and this has the effect of lowering the quantity of dollars demanded by euro zone countries, as well as lowering the quantity of euros supplied by euro zone countries. At the same time, cheaper euro zone exports in effect are cheaper US imports from euro zone countries, which increase the quantity of euros demanded by the US and increase the quantity of dollars supplied by US importers. As a result, US imports (debits) and euro zone exports (credits) increase, while US imports (credits) and euro zone exports (credits) decrease. Both current account imbalances are eliminated.

that a *zero balance in the balance of payments means there is a balance between the demand for and supply of a currency*.

The financial account and exchange rates

It was noted in Chapter 16 that most economists believe that what happens in the financial account is at least in part a reflection of developments in the current account. However, the financial account is also able to exert an important influence on the exchange rate. Suppose for example that Country X with a financial account surplus (and a corresponding current account deficit) is experiencing high rates of inflation, leading the central bank to pursue contractionary monetary policy by raising interest rates. This will attract an inflow of financial capital into Country X, and hence additional credits in the financial account corresponding to increased currency demand. In addition, there would be a fall in currency supply or decreased debits in the financial account as some domestic investors now prefer financial investments in the domestic market. More credits together with fewer debits in the financial account result in an excess of credits over debits, or an excess of currency demand over supply. Assuming that the current account deficit remains unchanged, the excess credits will cause an appreciation of the currency of Country X.

Similarly, an excess of debits over credits in the event of an outflow of financial capital would cause a depreciation of the currency. In the event of outflows due to currency speculation in expectation of a depreciation, there could be very large outflows of funds (known as **capital flight** to be discussed in Chapter 19, Section 19.2) leading to very significant exchange rate depreciation.

17.2 Comparing and contrasting exchange rate systems

LEARNING OBJECTIVES

In Chapter 16, Section 16.3, the last learning objective was to evaluate fixed versus floating exchange rate systems. Now that we have studied the connection between the balance of payments and exchange rates, we are in a position to evaluate exchange rate systems (AO3).

Degree of certainty for stakeholders

Fixed exchange rates

Under fixed exchange rates, there is a high degree of certainty for firms, consumers and the government because they know what exchange rates will be in the future. This certainty makes it easier for businesses to plan future investments domestically and abroad, sales of their products to other countries (exports), costs of imported inputs, and other activities because they do not have to take into account possible exchange rate changes that would change relative prices from country to country. Consumers can better plan travel abroad, purchases of imported goods and services and financial investments in other countries. Governments can similarly plan activities involving foreign transactions

TEST YOUR UNDERSTANDING 17.1

1 Using Bopland as an example, and its national currency, the bople, draw exchange rate diagrams to show why

  a a deficit in the current account is likely to result in a downward pressure on the value of the currency, and

  b a surplus in the current account is likely to result in an upward pressure on the value of the currency.

2 Explain why

  a a surplus in the financial account is likely to result in an upward pressure on the value of the currency, and

  b a deficit in the financial account is likely to result in a downward pressure on the value of the currency.

3 Using the example of Bopland, explain how its balance of payments deficit could be corrected in the three exchange rate systems (floating, managed, fixed).

(purchases of imported goods and services, payments of interest and capital on foreign loans, etc.). As a result there may be a better allocation of resources.

For the same reasons, fixed exchange rates favour international trade. The absence of exchanges rate changes makes it possible to calculate accurately the prices of goods and services in different countries.

Floating exchange rates

Floating exchange rates cause uncertainty as stakeholders cannot be sure what the value of currencies will be in the future. This may have negative effects on trade and investment flows due to an inability to plan accurately for the future.

In addition, large and abrupt exchange rate changes can cause serious problems for countries that depend heavily on exports. Occasionally these can result in financial crises, due to very large current account deficits (which may sometimes require intervention by the International Monetary Fund; see Chapter 20) through loans that help finance the deficits.

The role of foreign currency reserves

Fixed exchange rates

Central bank intervention to maintain a fixed exchange rate (Chapter 16, Section 16.3) requires sufficient supplies of reserves of foreign currencies. If there is a current account deficit, reserve currencies can be sold to buy the domestic currency, thus creating credits in the financial account to offset the excess debits in the current account. Problems can arise if central banks do not have enough reserves to carry out the necessary interventions (see below).

Floating exchange rates

Under floating exchange rates there is no need for central banks to hold foreign currency reserves since there is no need for intervention in foreign exchange markets. The balance of payments balance entirely through market forces.

Correction of current account imbalances

Fixed exchange rates

Under fixed exchange rates, there are no easy methods to correct imbalances in the balance of payments. External shocks (such as a sudden increase in oil prices leading to current account deficits for oil importers) cannot be handled quickly and easily. Large or persistent current account deficits require large quantities of foreign currency reserves or access to foreign borrowing. If these are not readily available, the country must resort to contractionary policies, trade protection or exchange controls, all with serious negative repercussions (see below). If current account deficits persist, the country may have to devalue its currency.

Floating exchange rates

One of the most important advantages of flexible exchange rates is their ability to adjust automatically to excess demand or supply of the domestic currency, thus bringing about a balance in the balance of payments (see Section 17.1 above). A current account deficit is eliminated through currency depreciation; a surplus is eliminated by currency appreciation.

As a result there is easy adjustment to external shocks. A sudden increase in oil prices, for example, leading to a current account deficit is met by a fall in the value of the currency. This mechanism allows the economy to be shielded against the effects of negative external developments.

Effects on inflation

Fixed exchange rates

If a country has a rate of inflation that is higher than that of its trading partners, the foreign demand for exports falls as these are less competitive, while the demand for imports rises, and the country is likely to develop a current account deficit. However, there is no possibility of the current account deficit being corrected through depreciation. This means that countries with a fixed exchange rate are encouraged to use fiscal policy to maintain a low and stable rate of inflation in order to be able to maintain their export competitiveness. This may mean use of contractionary policy which creates recession.

Floating exchange rates

In the event that the rate of inflation is greater than that of its trading partners, so that the country develops a current account deficit, this can be corrected through currency depreciation. However, depreciation leads to higher costs of imports, which in turn can create cost-push inflation due to the higher cost of imported inputs.

Flexibility offered to policy-makers

Fixed exchange rates

Fixed exchange rates do not offer flexibility to policy-makers. The need to maintain the exchange rate at a fixed level forces the government and central bank to pursue a range of policies that come with certain disadvantages, particularly in the case where the currency is under pressure to lose its value. These policies were considered in Chapter 16, Section 16.3, where we saw that:

- interest rate increases attract financial investments but have contractionary effects in the domestic economy
- borrowing from abroad also increases inflows of funds, but extensive borrowing may lead to high levels of debt resulting in a range of problems (see Chapter 19)
- contractionary fiscal and monetary policies to limit imports may create a recession and unemployment in the domestic economy
- trade protection to limit imports results in increased inefficiency in production, increased domestic and global misallocation of resources and may result in retaliation.

Floating exchange rates

Floating exchange rates offer greater flexibility to policy-makers. Domestic economic policy does not need to respond to balance of payments problems, and can be carried out in accordance with domestic priorities. For example, if there is a current account deficit, there is no need to pursue contractionary fiscal and monetary policies (which would create a recession). The government can pursue expansionary fiscal and monetary policies, and the current account deficit will automatically be corrected through currency depreciation.

Effects on speculation

Fixed exchange rates

Speculation is limited since the exchange rate cannot move up or down. Therefore fixed exchange rates remove a cause of currency instability due to speculation. (An exception occurs when people believe that a country may devalue or revalue its currency, in which case there is room for speculation.)

Floating exchange rates

Currency speculation which occurs under floating exchange rates can be destabilising. If speculators expect a currency to depreciate due to large current account deficits, they can sell the currency in anticipation of its depreciation. As a result they cause it to depreciate more than it otherwise would.

Evaluating managed exchange rates (Supplementary material)

The managed float, currently in use today, came about spontaneously following the collapse of the system of fixed exchange rates in 1973. Supporters of the managed float argue that it is superior to fixed exchange rates because it offers flexibility to pursue policies according to the needs of the domestic economy. Its flexibility further allows economies to adjust more easily to shocks (such as abrupt increases in the price of oil). Also, they claim it is superior to floating exchange rates because it offers governments the opportunity to prevent very sudden and large exchange rate fluctuations, and it also works to make currency speculation more difficult because speculators do not know if and when a central bank will intervene in a currency market to change the value of a currency.

Critics of the managed float argue it cannot do enough to prevent large currency fluctuations, which are especially damaging to economies highly dependent on exports. They argue that sharp drops in exchange rates played a major role in bringing about the severe financial crisis experienced by some Asian countries in the late 1990s. Also, the managed float does not appear to be successful in eliminating large trade imbalances, as the experience of the United States indicates. Further it offers countries the opportunity to 'cheat' by undervaluing their currencies and gaining an unfair competitive advantage (the 'dirty float'; see Chapter 16).

Few economists today would suggest returning to a fixed exchange rate system; however, some economists emphasise the need for increased international collaboration that would exercise some oversight over the management of exchange rates.

TEST YOUR UNDERSTANDING 17.2

1. Discuss the advantages and disadvantages of fixed and floating exchange rate systems.
2. (Optional) Describe some arguments in favour of and against managed exchange rates.

17.3 Evaluating monetary union

LEARNING OBJECTIVES

In Chapter 15, Section 15.3, the last learning objective was to discuss advantages and disadvantages of monetary union. Now that we have studied exchange rates we can evaluate monetary union (AO3). (You are reminded to review this section in Chapter 15.)

In Chapter 15, we saw that *monetary union* occurs when the member countries of a common market adopt a common currency and a common central bank responsible for monetary policy, such as the European Monetary Union whose member countries have adopted the euro. It was briefly noted that in some ways monetary union is like a system of fixed exchange rates for the countries that participate in the union, but without any possibility of revaluing or devaluing their currencies. In other words, by adopting a single currency it is as though they permanently fix the values of their currencies against each other. From this perspective, some (though certainly not all) of the advantages and disadvantages of monetary union are similar to the advantages and disadvantages of fixed versus flexible exchange rates. But, additionally, monetary union has a further very significant characteristic: the creation of a single currency, overseen and controlled by a single central bank. This characteristic offers further potential benefits, as well as possible costs.

Advantages of monetary union

A single currency eliminates exchange rate risk and uncertainty. Exchange rate fluctuations create risks and uncertainties for traders and investors, who do not know what the future exchange rates will be. A single currency eliminates the risks and uncertainties, with benefits for importers and exporters, consumers and investors, thereby encouraging trade and investments across boundaries. This contributes to achieving a more efficient allocation of resources.

A single currency encourages price transparency. Price transparency refers to the ability of consumers and firms to compare prices in all the countries that have adopted a common currency without having to make exchange rate calculations and conversions. This makes it easier for all economic decision-makers to see price differences quickly and accurately across countries and make price comparisons, and has the effect of promoting competition and efficiency.

A single currency eliminates transaction costs. Whenever there is a conversion of one currency into another, the banks (or whatever institution performs the conversion) charge a fee for the conversion; this is a type of transaction cost. A single currency eliminates the transaction costs of the conversion, resulting in significant savings that have the effect of encouraging trade, investments and international financial flows of all kinds. It has been estimated that the savings from the elimination of transaction costs of currency conversions within the euro zone countries amounts to about 1% of the combined GDPs of the countries involved.

A single currency promotes a higher level of inward investment. Inward investment refers to investments from outsiders towards the member countries with a common currency, and these can be expected to rise because of the absence of currency risk within an expanded market, resulting in greater economic growth.

Low rates of inflation give rise to low interest rates, more investment, increased output. If a member country has a high rate of inflation, its exports become less competitive, possibly a current account deficit. But there is no possibility of currency depreciation to regain competitiveness since there is no domestic currency. Therefore member countries become committed to maintaining a low rate of inflation. Low rates of inflation contribute to creating important benefits: low rates of interest, higher levels of investment and higher levels of output. A number of euro zone countries, including southern European countries as well as economies that transitioned from central planning to market economies, had high rates of inflation prior to monetary union. All have brought down their inflation rates to low levels with adoption of the euro.

Disadvantages of monetary union

A single currency involves loss of domestic monetary policy as an instrument of economic policy. For all euro zone countries, monetary policy is the responsibility of the European Central Bank, the objective being price stability for the region as a whole. Each individual country, whatever its particular circumstances (higher or lower inflation, unemployment, etc., than the average of the euro zone countries), is unable to carry out its own monetary policy to influence the rate of interest and hence the level of economic activity within its boundaries.

Monetary policy pursued by the single central bank will impact differently on each member country, depending on its own particular circumstances. Since countries are likely to differ from each other with respect to degrees of inflation, unemployment, etc., the single monetary policy pursued by the single central bank is likely to have different impacts on each of the member countries.

A single currency involves loss of exchange rates as a mechanism for adjustment. If a member country has a trade deficit with another member country, it no longer has its own national currency that could depreciate (in a flexible exchange rate system) or devalue (in a pegged system) in order to correct the imbalance. Similarly, if a member country has a higher rate of inflation and its goods become less competitive abroad, the problem cannot be solved by currency depreciation or devaluation. Instead it has to use contractionary fiscal policy to correct the imbalance which leads to recession.

Fiscal policy is constrained by the convergence requirements. Whereas each member country retains the ability to carry out its own fiscal policy, there are certain restrictions imposed by *convergence requirements*. In the case of European Monetary Union, total public debt cannot be greater than 60% of GDP and the budget deficit of any given year cannot be greater than 3% of GDP. Whereas this is seen by many to be an advantage (because it promotes fiscal discipline), others view it as a restriction of the authority of the government of a country, and consider it to be a disadvantage of monetary integration. (Many EMU countries have at various times violated the convergence requirements.)

A single currency overseen by the single central bank involves loss of national governments' authority in economic policy-making. It is no longer national governments and national central banks that are responsible for economic policy, but rather the central bank that oversees the monetary system of all the member countries of the monetary union. Moreover, authority is transferred away from democratically elected national governments towards an independent body that may be unelected (as in the case of the European Central Bank).

Many economists argue that the EMU may not survive over the long term if it does not reform. One change they consider to be essential is *fiscal union*. Fiscal union is an even higher form of economic integration that would involve the establishment of a central fiscal body with authority over all member countries. This body would have the powers to tax the member countries through central taxation policies and would also spend tax revenues in accordance with centrally determined policies. Such a fiscal union has never been attempted to date among any group of countries. The countries of the euro zone each make decisions about taxing and spending on a national level.

Economists consider fiscal union to be essential to the long-term success of monetary union in order to deal with the obstacles posed by the elimination of independent monetary and exchange rate policies. (See Real world focus 17.1.) Examples noted are the United States, Germany and Switzerland, all comprised of states with their own state governments headed by an overarching federal (national) government with powers over all the states to tax and spend. This means that a depressed state, for example, which has no monetary policy of its own or currency that it can depreciate to increase exports, would receive financial assistance to emerge from its recession. However, there remain serious political constraints to the development of fiscal union among the euro zone countries. The reason for this lies mainly with the loss of sovereignty that many countries fear results from increasing economic integration.

TEST YOUR UNDERSTANDING 17.3

Discuss advantages and disadvantages of monetary union.

REAL WORLD FOCUS 17.1

The Greek crisis reveals the weaknesses of EMU

The problem of government debt and EMU

Figure 17.2: Athens, Greece. Pensioners rally against austerity policies and pension cuts

During the first decade since the establishment of the European Monetary Union (EMU), there were no major problems. But with the onset of the global financial crisis in 2008, there was concern that as EMU countries went into recession, the 'one-size-fits-all' monetary policy and the inability to pursue a national exchange rate policy could pose serious problems for countries with government deficits and debt. When a country goes into recession, the government deficit (and debt) increase because of falling tax revenues and increasing government expenditures. By 2009, many of the 16 euro zone countries had deficits and debt exceeding the 3% and 60% of GDP requirements of the EMU. The bombshell exploded when it was announced that the Greek public deficit and debt were so large that Greece was having difficulties borrowing to finance (pay for) its deficit and debt.

With Greece being a small country whose GDP amounts to only 2.6% of euro zone GDP, why should anyone worry about Greek debt? The answer had to do with the euro itself. Because several euro zone countries had large and rapidly growing budget deficits and debts, financial investors (investors in government bonds, or government debt) became worried that the borrowing countries might default, in which case their investments in bonds would be worthless. When this happens, it becomes more and more difficult for governments to borrow. The reason is that financial investors require much higher interest rates on bonds, in order to compensate them for the increased risk (this is called a 'risk premium'). If interest rates on bonds rise to very high levels, governments can no longer afford to borrow from financial markets. If they can no longer borrow, a serious risk of default arises, and one country could begin defaulting after another. Note there is a self-fulfilling prophesy at work here.

So the answer to the problem would appear to be that euro zone countries must lower their deficits and debts. How could they do this? This is where some of the difficulties of the euro presented themselves.

The importance of policies to promote growth

The easiest way to come out of a situation of very high debt is to 'grow' out of it. As Nobel Prize-winning economist Paul Krugman explains, in 1946, after the Second World War, the United States had a very large federal government debt of 122% of GDP. Yet investors were not concerned. Ten years later, debt as a share of GDP had been cut in half. In fact, the US government did not even 'pay back' its debt during that period; what happened is that GDP grew, roughly doubling over the same period.[2]

Therefore, what is needed in countries with large deficits and debts is economic growth. To achieve growth, countries need to rely on expansionary policies. Countries with high debts like Japan, the United Kingdom and the United States have their own currencies, which they can depreciate, giving a competitive advantage to their exports. They can also conduct an independent, expansionary monetary policy, which would further encourage growth. For example, the central banks of some countries (like Japan and the United States), in their efforts to fight recession, lowered their interest rate to nearly zero in the early years of the global financial crisis. The hope was that as their economies came out of recession and began to grow, they could gradually emerge from their situation of high debt.

[2] Paul Krugman, 'Learning from Greece' in the *New York Times*, 8 April 2010.

CONTINUED

Yet such independent exchange rate or monetary policy action is not an option for countries within the euro zone. Greece, as well as other southern European countries in the EMU, needed very low interest rates to encourage economic activity. However, the European Central Bank, influenced by Germany's concern about inflation, did not allow the interest rate to fall for fear this would create inflationary pressures. Therefore, for countries in the EMU that needed to reduce their deficits and debts, the only available domestic policy tools that could support growth were supply-side measures intended to increase international competitiveness.

These, however, need a long time to take effect. Another problem is that supply-side measures by themselves are not enough to solve the debt problem. Highly indebted governments must also try to deal directly with deficits and debt by cutting government expenditures and increasing taxes. This is the very opposite of what is required for growth, since these policies are actually contractionary: *they work to lower GDP rather than increase it*.

Why didn't Greece leave the EMU?

An obvious course of action would have been for Greece to leave the EMU, return to its national currency (the drachma), and regain its ability to conduct independent monetary and exchange rate policies. However, this option would not be seriously considered by Greece or any other euro zone country, except as a last resort. If one country were to leave the euro zone due to debt problems, this could lead to a chain of events with potentially disastrous consequences: it could shatter the confidence of investors in the ability of euro zone countries to pay back their debts, leading to sales of euro assets, a drastic fall in the value of the euro, inability of governments to borrow, a series of defaults in countries within the EMU, possibly extending to other indebted countries outside the euro zone (such as the United Kingdom), and a massive global financial crisis with a very deep global recession.

The importance of co-operation between euro zone countries

In a situation like this, an EMU-wide policy of mutual assistance and co-operation would have been very useful. The European Central Bank could have considered lower interest rates to help the countries in great need of some stimulus. The European Investment Bank could have considered investments that would counteract the effects of budget cuts and tax increases. A 'solidarity fund' could have been established to offer low-interest loans for investments in countries needing to encourage their growth. In fact, had there been an EMU-wide fiscal authority the problem might have been avoided altogether, or at least might have been far less severe. Prominent economists, including Joseph Stiglitz, Paul Krugman (both Nobel Prize winners) and Martin Feldstein, have compared the euro zone to the United States, explaining that the reason why the United States works well as a currency area is that there is a centralised fiscal policy. As Paul Krugman writes:

> *'Consider the often-made comparison between Greece and the state of California. Both are in deep fiscal trouble, both have a history of fiscal irresponsibility. . . .*
>
> *But California's fiscal woes just don't matter as much, even to its own residents, as those of Greece. Why? Because much of the money spent in California comes from Washington, not Sacramento. State funding may be slashed, but Medicare reimbursements, Social Security checks, and payments to defence contractors will keep on coming.*
>
> *What this means, among other things, is that California's budget woes won't keep the state from sharing in a broader US economic recovery. Greece's budget cuts, on the other hand, will have a strong depressing effect on an already depressed economy.'*[3]

In early 2010, the euro zone countries, under the leadership of Germany, failed to present a united front in the face of a member country's debt problems. According to Gustav A. Horn, Director of Germany's Macroeconomic Policy Institute, 'the EU should have explained credibly and clearly that it would take

[3] Paul Krugman, 'A money too far' in the *International Herald Tribune*, 6 May 2010.

CONTINUED

a shared common responsibility of an equal member of the common internal market'.[4] Instead, as the EU hesitated and delayed taking action, financial markets reacted by demanding ever-increasing and prohibitively high interest rates in order to lend to Greece, making it impossible for Greece to borrow from private investors. As a result, Greece was forced to borrow from the EU and the International Monetary Fund (IMF, to be discussed in Chapter 20). In Horn's words,

> *'As a result of the active assistance of the German government, what has happened is exactly the thing that was supposed to be avoided, and what could have been avoided – namely that Greece has indeed been forced to ask for financial help from other European countries.'*[5]

The EU/IMF loans were made on condition that Greece pursues highly contractionary policies, involving very harsh cuts in government spending and huge tax increases. Because of its large debt, Greece would in any case have had to pursue contractionary policies, as explained above. However, with the lack of an independent monetary and exchange rate policy, the budget cuts and tax increases had to be far greater. Since the Greek economy was already in recession, this had the effect of making the recession much deeper, at a high human cost in terms of unemployment, reduced social benefits and increasing poverty.

According to Joseph Stiglitz:

> *'. . . one hoped the Greek tragedy would convince policymakers that the euro cannot succeed without greater cooperation (including fiscal assistance). But Germany (and its Constitutional Court), partly following popular opinion, opposed giving Greece the help that it needs.*
>
> *To many, both in and outside of Greece, this stance was peculiar: billions had been spent saving big banks, but evidently saving a country of 11 million people was taboo.*[6]

The growing seriousness of the European debt crisis

By the spring of 2011, it was becoming increasingly apparent that the 'bailout' for Greece by the EU and IMF was based on a serious shortcoming: negative growth rates experienced by the Greek economy due to the severe recession made it much more difficult for Greece to make its debt repayments. Higher taxes and lower government spending led to lower incomes and lower tax revenues, and therefore an increased need to continue to borrow to make debt repayments. Greece was getting caught in an unsustainable debt situation, or 'debt trap', where a country must keep taking out more and more new loans to pay the old ones, thus increasing rather than reducing the level of its debt.

In the meantime, the European debt crisis was spreading to other countries. In November 2010, Ireland requested financial assistance from the EU and IMF due to the accumulation of government debt arising from efforts to rescue its failing banks. This was followed by a similar request in May 2011 by Portugal, that was also facing mounting debt problems due to an excess of government spending over tax revenues. By the summer of 2011, there were major fears that Spain and Italy might also be requiring bailouts.

The role of the multiplier in Greece's contraction

In fact, the contraction of the Greek economy was far greater than what had been expected. In 2013 the International Monetary Fund (IMF) made the astounding revelation that it had underestimated the size of Greece's multiplier (see Section 13.4, Chapter 13). This meant that the reductions in spending that were forced on the Greek economy by its lenders had a greater than expected effect on real GDP, causing a far deeper recession.*

* Source: https://www.nytimes.com/2013/06/06/business/global/imf-concedes-major-missteps-in-bailout-of-greece.html

4 Gustav A. Horn, 'How Germany made the Greek debt crisis worse', Speigelonline International, 27 April 2010.

5 ibid

6 Joseph Stiglitz, 'Reform the euro or bin it' in the *Guardian*, 5 May 2010.

CONTINUED

The crisis reveals the flawed nature of the EMU

The common currency of the euro zone countries means that exchange rate changes cannot correct current account deficits or surpluses. Some euro zone countries (especially Germany) have increased their export competitiveness, accumulating large trade surpluses, while others, including Greece, have lost competitiveness, and therefore have trade deficits. Germany's very large trade surpluses (about two-thirds of Germany's export revenues come from euro zone countries) owe their existence partly to an 'undervalued' currency. If the euro did not exist, its national currency would appreciate, it would lose part of its strong competitive advantage, and its growth performance would be weakened. At the same time, weaker euro zone countries' currencies would depreciate, giving them the competitive advantage they strongly need to support their economies. As the German Finance Minister has admitted, Germany became so prosperous 'because it has more advantages from European integration than any other country'.[7]

Since exchange rate adjustments are not possible within the euro zone, there is an urgent need for another adjustment mechanism for EMU to work. Many economists argue that the present EMU institutions must be supplemented by an EMU-wide fiscal authority, with powers to tax, spend and invest, transfer funds and monitor member countries. This would allow depressed areas to receive the necessary funds and investments to boost their growth. Yet there is tremendous opposition to this within the EMU countries themselves. Some of the countries that are by far the strongest opponents to close co-operation and fiscal policy co-ordination are the same ones that have gained the most from European integration.

The end of the bailout programmes

In August 2018, eight years after the first bailout by the IMF and EU, and after two additional bailouts in subsequent years, Greece returned to borrowing from financial markets. This was considered to be a great moment.

However, the costs for the Greek economy and people have been tremendous. Greece is just starting to come out of the greatest depression of any advanced country in recorded history. As a result of the punishing conditions imposed in exchange for the bailouts, real GDP is 25% smaller, while unemployment remains the highest in the euro zone countries. In 2013, youth unemployment peaked at nearly 60% and overall unemployment at nearly 28%. Pensions have dropped significantly and disposable incomes have dropped by one-third. One person in three lives at or below the poverty line. Throughout this period, highly optimistic projections for growth were consistently disproven and growth rates still remain far below those of other countries in Europe.

Over 400 000 Greeks, roughly 4% of the population, most of them in their twenties and thirties, left the country in search of work, with little likelihood of their ever returning in view of employment prospects in Greece. Most of them were highly skilled and educated. This is a massive drain of human capital.

The problem with austerity and budget surpluses

A key issue is that Greece has been forced by its creditors to run a primary budget surplus, which is an excess of government revenues over spending, not including interest payments. The reason for this is that a budget surplus is the only way that a borrowing country can repay its debts, and creditor countries like Germany wanted to be sure they get their money back.

However, running a budget surplus is by its nature *contractionary* fiscal policy, since it takes away from the economy's spending stream in the form of taxes more than it returns in the form of government spending. Therefore the only way this could work without sending the economy into a deep recession is if there is economic growth.

[7] Quoted in 'Forget Greece: Europe's real problem is Germany' in the *Washington Post*, 21 May 2010.

CONTINUED

Instead, with negative growth rates over several years, austerity and the necessity to accumulate budget surpluses, the Greek experience has shown the disastrous consequences of such a policy mix, which works to deepen the recession, making debt repayment come at a very high human cost. Nobel prize-winning economist Joseph Stiglitz notes the following:

> *'Germany of all countries should understand this. At the end of World War I, Germany was made to pay reparations by the Treaty of Versailles. To finance the reparations, it would have to run a surplus. Keynes[8] correctly predicted that German reparations and the resulting German surplus would cause a German recession or worse. The depression in Germany that followed had disastrous political consequences not just for Germany for the entire world . . . Given the history, it is shocking that Germany and [Greece's creditors] have demanded that Greece and other crisis countries maintain large primary budget surpluses.'[9]*

The fear is that if recession hits Greece again, there will be no more room for fiscal tightening, as government spending has no room for falling further while punitive taxes cannot rise further. Since 2015 there has been a difference of opinion between the IMF and European creditors. The IMF has been advocating lenient debt restructuring on the grounds that rapid debt reduction decreases economic activity, as a result making debt repayment far more difficult. The EU on the other hand has been rejecting this approach relying on harsher measures.

Over the longer term, it appears unlikely that the EMU can work without fiscal mechanisms of coordination, mutual support and surveillance, which would also allow its members to share both the benefits and costs of EMU membership more equally.

Sources: *Krugman 2010a; Krugman 2010b; Stiglitz 2010a; Stiglitz 2010b; Feldstein, 2010; 'Acropolis now', The Economist, 2010; http://time.com; www.economist.com*

Applying your skills

1 a Explain how economic growth helps countries lower their debt-to-GDP ratio.

  b Using an *AD-AS* diagram, explain why contractionary policies make it more difficult to lower the debt-to-GDP ratio.

2 Explain why a common currency can be interpreted as being 'undervalued' for countries with trade surpluses and 'overvalued' for countries with trade deficits.

3 Use a *PPC* diagram to show the effects of emigration of young people on Greece's future growth prospects.

4 Use a circular flow diagram and the concepts of leakages injections to show why running a budget surplus by the government leads to contraction of the economy.

5 Explain how Greece's experience illustrates the disadvantages of monetary union.

[8] This is a reference to John Maynard Keynes who was introduced in Chapter 1.

[9] Joseph E. Stiglitz, (2017) *The Euro and its threat to the future of Europe*, Penguin Books.

17.4 Understanding current account deficits and surpluses

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- evaluate the implications of a persistent current account deficit with regard to exchange rates, interest rates, sale of domestic assets to foreigners, debt, credit ratings, demand management, and economic growth (AO3)
- explain and evaluate methods to correct a persistent current account deficit: (AO2, AO3)
  - expenditure switching
  - expenditure reducing
  - supply-side policies
- explain and apply the Marshall-Lerner condition and J-curve effect (AO2)
- draw a diagram illustrating the J-curve referring to the Marshall-Lerner condition (AO4)
- evaluate the implications of a persistent current account surplus with regard to exchange rates, domestic consumption and investment, inflation, employment, and export competitiveness (AO3)

Consequences of persistent current account deficits

Most balance of payments problems usually arise in connection with current account deficits, mainly due to an excess of imports over exports over long periods of time. This allows countries to enjoy increased levels of consumption over what they produce, as we saw in the *PPC* diagram in Figure 16.7(a), however this cannot go on indefinitely. While current account deficits for short periods of time, or current account deficits that alternate with current account surpluses, do not generally pose problems, over the long term there are likely to be several negative consequences.

Current account deficits, as we know, are paid for by financial account surpluses. As the central bank does not have endless amounts of foreign currency reserves to pay for a current account deficit, it must do so either through loans (borrowing from abroad) or by selling some of its physical or financial assets; all these methods result in inflows of foreign exchange. Yet both borrowing and sales of assets may pose problems if pursued for a long time.

- **Depreciating exchange rate.** A current account deficit puts a downward pressure on the exchange rate. Large depreciations can lead to imported inflation. If there is a risk of default the downward pressure on the currency is much stronger because people don't want to hold currencies whose value is expected to fall further, and the currency becomes vulnerable to speculation.
- **Possible need for higher interest rates to attract foreign financial investments, leading to recession.** If a country has difficulty getting loans, it may have to increase its interest rates to attract financial investments. However, higher interest rates discourage domestic investment and consumption spending, possibly creating a recession in the economy.
- **Foreign ownership of domestic assets.** The need for inflows of funds, or credits, in the financial account may lead countries to sell domestic assets to foreigners, such as stocks in the stock market, real estate or factories, all of which eventually may lead to loss of control over its assets.
- **Increasing levels of debt.** If a country borrows over long periods of time, it runs the risk of accumulating so much debt that it may be unable to pay it back; this is called a risk of default. Risks of default, along with actual default, come with many problems, such as significant currency depreciation, difficulties of getting more loans and painful demand-side policies (see below).
- **Cost of paying interest on loans.** The interest payments that must be made on loans use up national income of the country that could have been spent elsewhere in the domestic economy, such as for investment or provision of merit and public goods.
- **Fewer imports of needed capital goods.** Interest payments on loans also use up scarce foreign exchange earnings (from exports) that could have been used on imports of capital goods or other inputs

for production; the country may therefore not be in a position to secure all needed imports for production.

- **Poor international credit ratings.** International agencies rank countries according to how 'credit-worthy' they are, meaning how likely they are to repay their loans in full and on time. This is called a **credit rating**. Countries with large and persistent current account deficits have low credit ratings, making it more difficult to get more loans in the future (no one wants to lend to a country that may be unable to pay back its loans). Under such circumstances, a country may have to raise its interest rates very high to attract foreign financial capital, and this can create a serious recession or make an existing recession deeper. In addition, the ability of a country to go on indefinitely financing its current account deficit by selling off its assets depends very much on the confidence that foreign investors have in the domestic economy and currency. If there is a belief that the currency may depreciate substantially, or that the economy will not perform well in the future, they may be unwilling to continue to invest in the country, or they may even try to sell their assets in the country, in which case the country will be unable to finance its current account deficit. This is likely to result in a significant and rapid depreciation of the domestic currency.
- **Painful demand management policies.** Countries with serious current account deficits must often pursue contractionary policies. Contractionary monetary and fiscal policies have the effect of lowering incomes, which in turn lead to lower imports that may help to reduce the current account deficit. We will consider these policies below
- **Possibility of lower economic growth.** If loans accumulate over long periods of time, the cumulative impacts of the above may mean lower economic growth, as resources are used up on interest payments and loan repayments.
- **Lower standard of living in the future.** In order to be able to pay back the loans in the future, the local population will at some point have to consume less than they produce, giving rise to a decline in their standard of living. The reason is that paying back their debts requires a financial account deficit, corresponding to a current account surplus. (This is similar to one's personal finances: if you spend more than you earn by borrowing, in the future when you pay back your debts you will have to spend less than you earn. (See also Figure 16.7(b) in Chapter 16). Countries that sell off their assets do not have to pay back loans; however, they are selling off a portion of their domestic property. In this case, they will have to consume less than the amount they produce in the future if they want to regain possession of their domestic assets.

However, these problems *need not necessarily arise*. Borrowing can lead to economic growth, and if *per capita* output and income increase, it becomes possible to have increased consumption of goods and services even as loans are being paid back. Important requirements for this to happen are:

- the current account deficit remains relatively small and does not get out of hand by excessive borrowing
- borrowed funds are used to finance imports of capital goods and other inputs needed in production (instead of consumer goods imports)
- some production is geared towards export industries so that exports increase, making increased export earnings possible (to help pay back loans and interest, and finance more capital goods imports).

Policies to correct persistent current account deficits and their evaluation

Expenditure reducing policies (reductions in aggregate demand)

Contractionary fiscal and monetary policies reduce aggregate demand and therefore output and incomes, leading to lower demand for imports. In addition, reduced aggregate demand is likely to give rise to a lower rate of inflation, making domestic goods more competitive, thus increasing exports. The combination of fewer imports and more exports may work to reduce the current account deficit. These policies are known as **expenditure reducing policies**, because they try to influence the levels of imports and exports by reducing domestic expenditures through lower aggregate demand.

This approach comes with disadvantages, as it may create a recession in the domestic economy. Moreover, there is also a risk that higher interest rates (contractionary monetary policy) leads to currency appreciation, which may discourage exports and encourage imports, partly cancelling out the beneficial effects of expenditure reducing policies on imports and exports.

Expenditure switching policies

Expenditure switching policies attempt to switch consumption away from imported goods and towards domestically produced goods.

Trade protection

Countries with a long-term current account deficit could resort to increased trade protection, creating or increasing barriers to trade. These policies can reduce the current account deficit by directly restricting imports; however, they have a number of negative effects, such as higher domestic prices of protected goods, lower domestic consumption, inefficiency and a domestic and global resource misallocation of resources. There also arises a danger that countries against which the protective barriers are imposed may retaliate with their own barriers, creating a spiral of protectionist policies with serious consequences on global trade and global growth.

Depreciation

The currency of a country with a persistent current account deficit is likely to face a strong downward pressure on its value. The government or central bank may allow the currency to depreciate, in which case it encourages exports (which become cheaper to foreigners) and discourages imports (which become more expensive to domestic buyers). This is another type of expenditure-switching policy, because it, too, switches consumption away from imports and towards domestically produced goods. However, this policy too may have negative effects on the domestic economy. Higher import prices due to the lower value of the currency often result in higher domestic inflation. If the imported goods include capital goods and other production inputs, firms experience higher costs of production, which may be passed on to consumers in the form of higher prices. This is a type of cost-push inflation, and by shifting the *SRAS* curve to the left, may have recessionary effects.

Supply-side policies to increase competitiveness

Market-oriented supply-side policies are intended to lower costs of production for firms, and by shifting the *SRAS* and *LRAS* curves to the right can result in lower rates of inflation (see Chapter 13, Section 13.6). Several policies, such as increasing competition, reducing the power of labour unions, reducing or eliminating the minimum wage, cutting business taxes, deregulation, and others, could have the effect of making firms more competitive in global markets. Over a long period of time, lower rates of inflation may increase exports, thereby addressing the current account deficit.

Countries can also use interventionist supply-side policies, such as support for training, education, research and development and industrial policies, to promote industries that produce for export.

A disadvantage of supply-side policies is that they generally take a long time to make their effects felt. For further evaluation of these policies see Chapter 13.

Marshall–Lerner condition and J-curve

Marshall–Lerner condition

When a country's currency devalues or depreciates, its imports become more expensive domestically and exports become less expensive to foreigners. This suggests that the quantity of imports decreases and the quantity of exports increases. The question of interest is whether a devaluation or depreciation reduces the size of a trade deficit (and therefore a current account deficit). The Marshall–Lerner condition is a condition that, if satisfied, allows devaluation or depreciation to lead to an improvement in a country's balance of trade (and therefore in its current account).

The Marshall–Lerner condition involves the devaluing/depreciating country's price elasticity of demand for imports (PED_m) and foreigners' price elasticity of demand for the country's exports (PED_x). Elasticities are important because what matters is not changes in the *quantities* of imports and exports, but rather changes in the *values* of imports and exports.

If *PED* for imports is less than one, a percentage increase in the price of imports (due to the devaluation/depreciation) leads to a smaller percentage decrease in quantity of imports demanded, so that the value of imports increases, producing a negative effect on the trade balance; if there is a trade deficit, this will tend to become larger. On the other hand, if the *PED* for imports is greater than one, the value of imports will fall, producing a positive effect on the trade balance. In general, the larger the *PED* for imports, the greater the scope for improvement in a trade deficit.

The PED for exports is also important. The larger the PED for exports, the larger the increase in quantity of exports that results following a devaluation/depreciation, and the larger the positive effect on the trade balance.

Yet it is not necessary for both PEDx and PEDm to be larger than one for a devaluation or depreciation to result in a smaller trade deficit. According to the

Marshall-Lerner condition, it is only necessary that their sum be greater than one.[10]

The Marshall-Lerner condition states the conditions under which a devaluation/depreciation will result in a smaller trade deficit:

The **Marshall–Lerner condition** states the following:

- If the sum of the *PED*s for imports and exports is greater than 1, i.e. PEDm + PEDx >1, devaluation/depreciation will improve the trade balance (will make a trade deficit smaller).
- If the sum of the two *PED*s is less than 1, devaluation/depreciation will worsen the trade balance (will make a trade deficit bigger).
- If the sum of the two *PED*s is equal to 1, devaluation/depreciation will leave the trade balance unchanged.

The greater the price elasticities of demand for imports and for exports, the greater the scope for improvements in the trade balance. With low *PED*s, it would be necessary to have large devaluations/depreciations to obtain significant trade balance improvements. The higher the *PED*s, the smaller the devaluation or depreciation needed to obtain trade balance improvements.

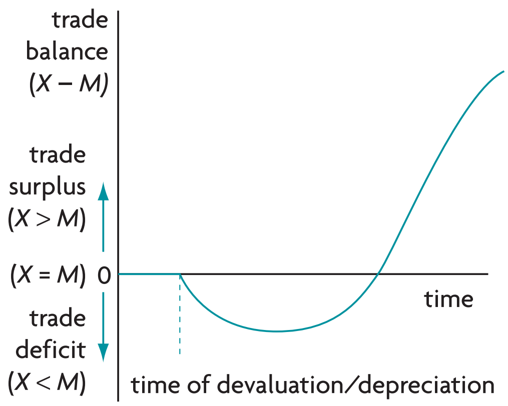
Understanding the Marshall-Lerner condition in more detail (Supplementary material)

If you would like to understand the Marshall-Lerner condition in more detail you may do so in the digital coursebook.

J-curve effect

A devaluing/depreciating country may see a worsening trade balance in the period immediately following the devaluation or depreciation of its currency; later, the trade deficit will begin to shrink, and the trade balance will begin to improve, provided the Marshall–Lerner condition holds. This is known as the **J-curve effect**, shown in a graph that plots the balance of trade (the value of exports minus imports) on the vertical axis and time on the horizontal axis, shown in Figure 17.3. All values greater than zero on the vertical axis illustrate a trade surplus, and all values less than zero illustrate a trade deficit. When the trade balance is equal to zero (at the origin), exports are equal to imports. It follows that as the values increase along the vertical axis, a trade deficit becomes smaller until it reaches $(X - M) = 0$, and above that becomes a trade surplus.

a Value of exports is equal to value of imports at time of devaluation/depreciation



b Trade deficit at time of devaluation/depreciation

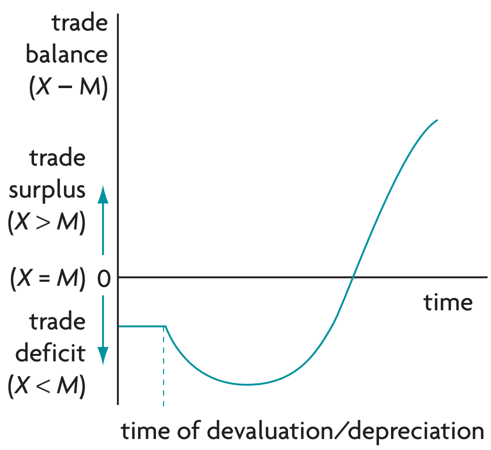


Figure 17.3: J-curve effect

[10] The PED for exports works differently than the PED for imports. Following a depreciation, the price of exports falls in terms of foreign currencies, but it remains the same in terms of the domestic currency. This means that following a devaluation/depreciation, the value of exports will *always* increase regardless of the PED of exports, because the value of exports is determined by multiplying a constant price (in terms of the domestic currency) by a larger quantity. If you are interested in finding out more about this see 'Understanding the Marshall-Lerner condition in more detail' presented as Supplementary material in the digital coursebook.

In part (a), the country initially has a value of exports equal to the value of imports (no trade deficit or surplus); following the devaluation/depreciation a trade deficit emerges which later becomes a trade surplus. In part (b), the country initially has a trade deficit, which becomes larger immediately following the devaluation/depreciation, and which then improves, eventually becoming a surplus. This is called a 'J' curve because of its shape.

The explanation behind the shape of the J-curve can be traced directly to the Marshall-Lerner condition. In the period immediately following a currency devaluation/depreciation, price elasticities of demand for imports and exports are very low, and the Marshall-Lerner condition is not satisfied: $PED_m + PED_x < 1$. Therefore the trade balance deteriorates. As time passes, *PED*s for imports and exports increase, and if there comes a point when $PED_m + PED_x > 1$, the trade balance begins to improve.

The reason for initial low *PED*s for imports and exports lies in time lags (time delays) between devaluation/depreciation and its effects on quantities of exports and imports demanded. Devaluation/depreciation involve changes in relative prices of imports and exports. However, consumers and producers need time to adjust to the price changes. Although imports have become more expensive domestically, they are still purchased for a variety of reasons, such as the time needed for buyers to become aware of the price changes, or prior commitments, or the time needed to place new orders, or particular preferences of buyers that need time to change. In the meantime, the price of exports has fallen for foreigners; however, they too have prior commitments, particular preferences and so on, therefore the quantity demanded increases only slowly. With $PED_m + PED_x < 1$ initially, the trade balance deteriorates, resulting in the downward-sloping portion of the J-curve.

As time passes, consumers and producers adjust to the changes in prices, *PED*s of both imports and exports increase, quantity of imports demanded falls and quantity of exports demanded increases and the balance of trade begins to improve (the upward-sloping portion of the J-curve).

Empirical evidence supports the existence of a J-curve. According to studies estimating *PED*s for imports and exports in developed countries, over short periods of time (less than six months) most manufactured goods have *PED*s that are too low to satisfy the Marshall–Lerner condition; the sum of *PED*s for imports and exports is less than 1, indicating the downward-sloping part of the J-curve (a worsening trade balance). In a period of more than six months and less than a year, the *PED*s for most products have increased to the point that the Marshall–Lerner condition is satisfied, so that the sum of the two *PED*s is greater than 1, leading to an improving trade balance.

Consequences of persistent current account surpluses

Countries with a current account surplus most likely have a trade surplus, with exports greater than imports. They are therefore net purchasers of assets abroad or net lenders to other countries (thus having financial account deficits). Persistent current account surpluses may lead to the following problems:

- **Low domestic consumption.** Large and persistent current account surpluses mean lower consumption levels and lower standards of living for the population since overall production is greater than consumption. This is clear from Figure 16.7(b) showing that counties with a trade surplus consume inside their *PPC*.
- **Insufficient domestic investment.** The financial account deficit (corresponding to the current account surplus) means that funds are leaving the country, resulting in a risk of insufficient domestic investment, limiting economic growth prospects.
- **Appreciation of the domestic currency.** A current account surplus puts an upward pressure on the value of a currency, which can lead to lower exports and higher imports (reduced net exports). This could lower the rate of growth of the domestic economy as a result of lower aggregate demand.
- **Inflation.** Lower aggregate demand due to reduced net exports puts a downward pressure on demand-pull inflation. Further lower import prices put a downward pressure on cost-push inflation
- **Employment.** The effects may be mixed. Lower aggregate demand may lead to higher unemployment as workers begin to lose their jobs. However unemployment may decrease in firms that enjoy lower import costs due to the appreciation.
- **Reduced export competitiveness.** As the domestic currency appreciates, exports become more expensive to foreigners, and this makes it more difficult for domestic firms to compete with firms abroad
- **Possibility of retaliation by trading partners through trade barriers.** Current account surpluses in some countries correspond to current account deficits in other countries. Current account surpluses persisting over long periods may prompt the deficit countries to impose trade restrictions to reduce their imports from the surplus countries. (See Real world focus 15.2.)

REAL WORLD FOCUS 17.2

Pakistan's current account deficit

Figure 17.4: Lahore, Pakistan. Freight train with cargo containers

In the period 2014–2017, the State Bank of Pakistan (SBP, Pakistan's central bank) managed the Pakistani rupee, preventing a depreciation that would most likely have occurred if the rupee had been left to free market forces. The depreciation is likely to have occurred because of a large trade and current account deficit in Pakistan.

The SBP used foreign exchange reserves to manage the rupee. Its interventions in the foreign exchange market resulted in a relatively stable rupee to US dollar exchange rate. But as the rupee could not fall to its free market level it became overvalued. As a result, it succeeded in bringing down the rate of inflation from 8.6% in 2014 to 2.9% in 2016, though this rose to 4.2% in 2017. However, the generally lower rate of inflation in relation to 2014 came at the cost of a larger trade and current account deficit. The trade deficit increased by more than 50% in the period 2014–2017 as the overvalued rupee resulted in more imports and fewer exports.

An alternative to using foreign exchange reserves to prevent rupee depreciation would have been for the SBP to increase interest rates. However, interest rates were maintained at low levels during this period.

In addition to foreign exchange reserves, the rupee was supported by Pakistani borrowing from abroad. From the end of 2014 to the end of 2017 Pakistan's external debt increased by nearly 30%.

At the end of 2017 the decision was made to allow the rupee to float (to be determined by market forces). The result was an immediate and very significant rupee depreciation. Yet even as the rupee fell, the trade and current account deficits continued to widen.

The rate of inflation in the meantime started to rise again. The SBP has increased interest rates. There are concerns that savings will be directed toward portfolio investments and savings deposits rather than to firms interested in investing in productive capacity. Confidence in the economy is falling. These factors are raising concerns about the future of the Pakistani economy. It is expected that economic growth in 2018–2019 will be the lowest in nine years.

Sources: *Hussain H Zaidi, 'Intervention indispensable to control rupee's value', The Express Tribune, 28 December 2018;*
N H Zuberi, 'PKR devaluation, policy rate to have devasting impact on economy', 4 December 2018;
Business Recorder; 'Pakistan's economic growth likely to touch nine year low', 18 December 2018;
https://dunyanews.tv/en/

Applying your skills

1 Use an exchange rate diagram to explain how the SBP succeeded in maintaining a stable value of the rupee in relation to the US dollar.

2 Explain how overvaluation of the rupee contributed to

  a lowering the rate of inflation, and

  b a widening trade deficit.

CONTINUED

3 The text states that an alternative to using foreign exchange reserves to maintain the value of the rupee would have been for the SBP to increase interest rates.

   a Using an *AD-AS* diagram, identify what kind of monetary policy this is and show the likely effects of rising interest rates on the Pakistani economy.

   b Using an exchange rate diagram explain the likely impact of this policy on the value of the rupee.

4 Referring to the concepts of debits and credits in the balance of payments, explain how two factors contributed to creating a zero balance in Pakistan's balance of payments.

5 Using a *PPC* diagram draw two points illustrating

   a where Pakistan could possibly be producing, and

   b where Pakistan could possibly be consuming.

6 Using an appropriate diagram explain one possible reason that accounts for the widening trade deficit in Pakistan after the depreciation of the rupee.

7 Using an appropriate diagram explain why the value of the rupee would have fallen if the SBP had not intervened.

8 Discuss the implications of Pakistan's current account deficit for the Pakistani economy.

9 Evaluate alternative methods that the SBP of government of Pakistan can pursue to correct the current account deficit.

Fiscal and monetary policy and conflicting objectives in an open economy (Supplementary material)

When an economy is open to international trade and financial flows, in addition to the goals of price stability, full employment and economic growth, it has the further goals of achieving a reasonable balance of trade and avoiding sharp fluctuations in its exchange rate. However, economies may be unable to achieve all these objectives at the same time. Governments often find that by pursuing a policy to correct one problem, they may create a problem elsewhere. If you are interested in reading about these conflicting objectives, you may do so in the digital coursebook.

TEST YOUR UNDERSTANDING 17.4

1 Discuss some problems faced by countries that have

   a a current account deficit and a financial account surplus, and

   b a current account surplus and a financial account deficit.

2 Discuss advantages and disadvantages of

   a expenditure reducing,

   b expenditure switching policies, and

   c supply-side policies as methods to correct a persistent current account deficit.

3 Explain why the price elasticities of demand (*PEDs*) for imports and exports are important in determining what will happen to a country's trade balance following a devaluation or depreciation of the domestic currency.

CONTINUED

4 Suppose a country has a deficit in its balance of trade, and depreciates or devalues its currency. Explain what will happen to its trade balance if

a its *PED* for imports is 0.2 and its *PED* for exports is 0.5,

b its *PED* for imports is 0.7 and its *PED* for exports is 0.3, and

c its *PED* for imports is 0.8 and its *PED* for exports is 0.6.

5 State the Marshall–Lerner condition.

6 Suppose that a country's *PED* for imports is 0.6 and for its exports is 0.2.

a Explain why a depreciation of its currency may initially lead to a larger trade deficit, and

b draw a diagram to show the effects.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Identify and research a country that has a serious and persistent current account deficit. How did this deficit come about? Try to identify some of the consequences of this deficit in the domestic economy. Is the government or central bank taking measures to reduce this deficit; if so what are these measures? Can you comment on the possible positive and negative consequences of these measures?

2 Identify and research a country with a persistent current account deficit that uses or has used expenditure switching or expenditure reducing policies. Discuss the consequences of these policies.

3 Identify one or more countries that manage their exchange rate and research the causes and consequences of exchange rate management.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Real world issue 2: Why do economies develop unevenly?

CONCEPTUAL UNDERSTANDINGS

1. The perceived meaning of development and ***equity*** change over time, with variations across cultures.
2. The objectives of ***economic well-being*** and ***equity*** in societies are pursued through interventions of governments and other economic agents.
3. The objective of ***sustainability*** is subject to a number of constraints.
4. Effective strategies should be based on consideration of the relevant economic, social and political contexts.

Development economics is a special area of economics that draws on and applies the principles of economics (micro, macro as well as international economics) to the problems of developing countries. However, it is much more than a simple application of these principles. The important questions we want to answer, such as what makes countries grow and develop, what prevents them from growing and developing, and what are the best ways to pursue growth and development objectives, cannot always be answered by use of economic principles alone. In addition, economic development depends on institutional factors, as well as on value judgements about what is considered to be good for development. Because of the special circumstances that developing economies find themselves in, it is often necessary to devise new ways of thinking about their economic problems, in order to understand and explain the nature of these problems, as well as to determine how they can best be addressed.

Rwanda. Farmer with a compost heap. Compost is recycled organic matter, usually waste products, made into organic fertilizer for agriculture, promoting sustainable development

Chapter 18

Understanding economic development

BEFORE YOU START

- What do you think makes for a decent standard of living for a country's citizens?
- What kind of indicators would you consider to measure standards of living and well-being?

This chapter examines the goals of sustainable development as well as the relationship between economic growth and economic development. In addition, it focuses on the methods that can be used to measure economic development, which is a process with many dimensions. We will discover that there are numerous ways to approach its measurement, each with advantages and drawbacks.

Photo: Niamey, Niger. A boy collects junk and garbage to sell locally

18.1 Sustainable development

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the meaning of sustainable development (AO2)
- explain the sustainable development goals (AO2)
- explain the relationship between sustainability and poverty (HL only) (AO2)

The meaning of sustainable development

The concept of sustainability, introduced in Chapter 1, concerns the joint preservation of the environment and the economy: how to achieve environmental preservation along with preservation of humankind's ability to provide goods and services to satisfy needs and wants into the future.

The problem of sustainability arises because of conflicts between environmental and economic goals. Economic goals involve efforts to increase the quantities of output produced and consumed; but focusing on economic goals while disregarding the environment may result in its irreversible destruction. Environmental goals involve the preservation of the environment; but focusing on environmental goals while disregarding the economy may result in humankind's inability to satisfy needs and wants.

The important question, then, is how to strike a balance between environmental and economic goals, so that both can be satisfied into the future. The answer to this question is provided by the concept of **sustainable development** (introduced in Chapter 1), defined as '*development that meets the needs of the present without compromising the ability of future generations to meet their own needs*'.[1] This concept was coined by the Brundtland Commission on Environment and Development, set up by the United Nations in the 1980s and named after the Norwegian Gro Harlem Brundtland who headed the Commission. It means that societies should pursue economic growth *that does not deplete or degrade natural resources*, so that future generations will not have fewer or lower-quality natural resources to satisfy their own needs.

Sustainable development goals

The **Sustainable Development Goals (SDGs)** are a set of seventeen goals that were developed at the United Nations Conference on Sustainable Development in Rio de Janeiro in 2012. In the words of the United Nations, 'The objective was to produce a set of universal goals that meet the urgent environmental, political and economic challenges facing our world.'[2]

The SDGs continue and expand upon the work that had begun years earlier by the Millennium Development Goals (MDGs), which ran until 2015. Like the MDGs, the SDGs are accompanied by numerous targets that are intended to be met within the 15-year period 2015 to 2030. The targets have one to three indicators used to monitor and measure countries' progress toward achieving the goals and targets. Indicators will be discussed below.

The SDGs with their corresponding targets and indicators are very important tools used by international organisations and national governments in their fight against poverty and efforts to achieve sustainable economic development. They are used systematically to monitor and measure progress (or setbacks) achieved in each country with respect to each of the goals.

The Sustainable Development Goals are listed in Table 18.1. If you are interested in more information such as the targets and indicators that are included for each goal you may visit their website by searching for 'UN sustainable development goals'.

[1] Brundtland Commission (World Commission on Environment and Development) (1987) *Our Common Future*, Oxford University Press.

[2] https://www.undp.org/content/undp/en/home/sustainable-development-goals/background.html

The relationship between sustainability and poverty (HL only)

We often think of environmental degradation as the by-product of high-income production and consumption activities resulting from increasing quantities of output produced and consumed (economic growth). This type of environmental damage has been termed 'pollution of affluence' and arises mainly from industrial production based on use of fossil fuels (such as oil) and using up common pool resources like clean air, rivers, lakes, and so on, leading to climate change.

However, there is another type of very important environmental damage, occuring mainly in developing countries, and arising from production and consumption activities that are due to poverty. This second type of environmental damage has been termed 'pollution of poverty', and is due to economic activities pursued by very poor people in an effort to survive.

According to the Brundtland Commission, which coined the term 'sustainable development', poverty is a cause of environmental destruction due to the overexploitation by poor people of their scarce environmental resources. Poor people lack modern agricultural inputs, and being too poor to buy inputs that preserve the soil's fertility, they deplete the soil's natural minerals, making soils less productive. Poor people usually have higher birth rates and higher population growth, creating pressures for them to open up new lands for agriculture. With suitable agricultural land becoming increasingly scarce, they cut down forests (deforestation) in search of new farmland, they move to fragile lands in mountains and hills, causing soil erosion, and they overgraze animals on pasture lands, depleting the nutrients there as well. Lacking modern energy sources, they also cut down forests to obtain firewood. Poor people have limited abilities to borrow to finance the purchase of inputs, and this works against their ability to make improvements in sanitation, irrigation, improved agricultural inputs and land improvements, which would reverse or reduce these types of environmental degradation.

The production and consumption activities of very poor people that endanger the environment and sustainability can also be interpreted as negative externalities involving overuse of common pool resources. In Figure 5.3 (Chapter 5), the *MPC* curve may be a farmer's private costs of farming, with the difference between the *MPC* and *MSC* curves representing overuse of forests that have been cleared for agriculture, or the overuse of soil leading to depletion of nutrients.

Whereas the pollution of poverty occurs mainly in developing countries, this is not to say that developing countries are not guilty of creating some 'pollution of affluence'. Increasingly, the pollution of affluence arises also in developing countries that grow by engaging in industrial production and consumption activities without regard for the environment.

TEST YOUR UNDERSTANDING 18.1

1 Explain

 a the meaning of sustainable development, and

 b the objectives of the Sustainable Development Goals.

2 (HL only) Explain the relationship between sustainability and poverty.

18.2 Measuring development

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain the multidimensional nature of economic development (AO2)
- explain the use of single indicators, including: (AO2)
  - GDP/GNI *per capita* at US$PPP
  - health and education indicators
  - economic/social inequality indicators
  - energy indicators
  - environmental indicators

Goal	Excerpts from the United Nations SDG site[3]
1 End poverty in all its forms everywhere	'Poverty is more than the lack of income and resources to ensure a sustainable livelihood. Its manifestations include hunger and malnutrition, limited access to education and other basic services, social discrimination and exclusion as well as the lack of participation in decision-making'
2 End hunger, achieve food security and improved nutrition and promote sustainable agriculture	'Right now, our soils, freshwater, oceans, forests and biodiversity are being rapidly degraded. Climate change is putting even more pressure on the resources we depend on, increasing risks associated with disasters, such as droughts and floods' 'A profound change of the global food and agriculture system is needed if we are to nourish the 815 million people who are hungry today and the additional 2 billion people expected to be undernourished by 2050'
3 Ensure healthy lives and promote well-being for all at all ages	'Many more efforts are needed to fully eradicate a wide range of diseases and address many different persistent and emerging health issues. By focusing on providing more efficient funding of health systems, improved sanitation and hygiene, increased access to physicians and more tips on ways to reduce ambient pollution, significant progress can be made in helping to save the lives of millions'
4 Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all	'The reasons for lack of quality education are due to lack of adequately trained teachers, poor conditions of schools and equity issues related to opportunities provided to rural children. For quality education to be provided to the children of impoverished families, investment is needed in educational scholarships, teacher training workshops, school building and improvement of water and electricity access to schools'
5 Achieve gender equality and empower all women and girls	'Gender equality is not only a fundamental human right, but a necessary foundation for a peaceful, prosperous and sustainable world. Providing women and girls with equal access to education, health care, decent work, and representation in political and economic decision-making processes will fuel sustainable economies and benefit societies and humanity at large'
6 Ensure availability and sustainable management of water and sanitation for all	'Clean, accessible water for all is an essential part of the world we want to live in and there is sufficient fresh water on the planet to achieve this.' 'Water scarcity, poor water quality and inadequate sanitation negatively impact food security, livelihood choices and educational opportunities for poor families across the world'
7 Ensure access to affordable, reliable, sustainable and modern energy for all	'Focusing on universal access to energy, increased energy efficiency and the increased use of renewable energy through new economic and job opportunities is crucial to creating more sustainable and inclusive communities and resilience to environmental issues like climate change'
8 Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	'Sustainable economic growth will require societies to create the conditions that allow people to have quality jobs that stimulate the economy while not harming the environment. Job opportunities and decent working conditions are also required for the whole working age population'
9 Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation	'Investments in infrastructure – transport, irrigation, energy and information and communication technology – are crucial to achieving sustainable development and empowering communities in many countries' 'Technological progress is the foundation of efforts to achieve environmental objectives, such as increased resource and energy-efficiency'

[3] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

Goal	Excerpts from the United Nations SDG site[3]
10 Reduce inequality within and among countries	'There is growing consensus that economic growth is not sufficient to reduce poverty if it is not inclusive and if it does not involve the three dimensions of sustainable development – economic, social and environmental' 'To reduce inequality, policies should be universal in principle, paying attention to the needs of disadvantaged and marginalised populations'
11 Make cities and human settlements inclusive, safe, resilient and sustainable	'Many challenges exist to maintaining cities in a way that continues to create jobs and prosperity without straining land and resources. Common urban challenges include congestion, lack of funds to provide basic services, a shortage of adequate housing, declining infrastructure and rising air pollution within cities'
12 Ensure sustainable consumption and production patterns	'Sustainable consumption and production is about promoting resource and energy efficiency, sustainable infrastructure, and providing access to basic services, green and decent jobs and a better quality of life for all' 'Since sustainable consumption and production aims at "doing more and better with less," net welfare gains from economic activities can increase by reducing resource use, degradation and pollution along the whole life cycle, while increasing quality of life'
13 Take urgent action to combat climate change and its impacts	'Climate change is now affecting every country on every continent Affordable, scalable solutions are now available to enable countries to leapfrog to cleaner, more resilient economies' 'Climate change, however, is a global challenge that does not respect national borders. It is an issue that requires solutions that need to be coordinated at the international level to help developing countries move toward a low-carbon economy'
14 Conserve and sustainably use the oceans, seas and marine resources for sustainable development	'Our rainwater, drinking water, weather, climate, coastlines, much of our food, and even the oxygen in the air we breathe, are all ultimately provided and regulated by the sea' 'Careful management of this essential global resource is a key feature of a sustainable future'
15 Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss	'Forests cover 30.7 per cent of the Earth's surface and, in addition to providing food security and shelter, they are key to combating climate change, protecting biodiversity and the homes of the indigenous population. By protecting forests, we will also be able to strengthen natural resource management and increase land productivity'
16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels	'The threats of international homicide, violence against children, human trafficking and sexual violence are important to address to promote peaceful and inclusive societies for sustainable development To tackle these challenges and build more peaceful, inclusive societies, there needs to be more efficient and transparent regulations put in place and comprehensive, realistic government budgets'
17 Strengthen the means of implementation and revitalise the global partnership for sustainable development	'A successful sustainable development agenda requires partnerships between governments, the private sector and civil society. These inclusive partnerships built upon principles and values, a shared vision, and shared goals that place people and the planet at the centre, are needed at the global, regional, national and local level'

Source: *Sustainable Development Goals https://sustainabledevelopment.un.org*

Table 18.1: Sustainable Development Goals

CONTINUED

- explain the use of composite indicators including: (AO2)
  - Human Development Index (HDI)
  - Inequality-adjusted Human Development Index (IHDI)
  - Gender Inequality Index (GII)
  - Happy Planet Index
- discuss strengths and limitation of the various approaches to measurement of economic development (AO3)
- discuss the relationship between economic growth and economic development (AO3)

The multidimensional nature of economic development

Economic growth versus economic development

Economic growth refers to increases in output (real GDP) and incomes over time, often measured on a *per capita* basis. *Economic development* refers to a process that leads to improved standards of living for a population as a whole. Increasing levels of output and incomes resulting from economic growth mean that societies can better satisfy the needs and wants of their populations and secure improvements in their standards of living. However, economic growth does not by itself guarantee that this will occur. Persisting poverty and the failure of many countries to secure long-lasting improvements in well-being, even with respectable rates of growth over extended periods of time, have shown that economic development is a highly complex and sometimes elusive process.

Why economic development is multidimensional

The evolving meaning of economic development

The meaning of economic development has changed over the years. In the 1950s and 1960s, economists thought that economic growth and economic development were practically one and the same. A famous development economist, Charles P. Kindleberger, wrote in 1965, 'Growth and development are often used synonymously in economic discussion, and this is entirely appropriate.'[4] It was believed that economic growth over long periods would automatically provide economic and social benefits for the entire population. Larger quantities of goods and services, including health care and education, and employment opportunities and social change would eventually be spread out over most people in an economy. This was termed the 'trickle-down theory': benefits of growth would eventually trickle down to everyone.

By the late 1960s and early 1970s, it was becoming clear that many developing countries were not performing according to expectations. The GNI *per capita* gap between rich and poor countries had more than doubled on average in the period 1950–1975.[5] While some less developed countries were growing rapidly (especially some oil-rich countries), others were experiencing very low or even negative growth rates (especially in Africa). The number of people living in extreme poverty (defined then as living on less than US$1 per day) was increasing rather than decreasing. It was apparent that the benefits of economic growth were not 'trickling down' to the poorest members of society.

Economists began to understand that what was needed was an approach that would directly deal with the problems of developing countries, and specifically the problem of persisting poverty. A new perspective emerged emphasising redistribution of income and wealth, and improved access of the poor to basic goods and services.

The many dimensions of economic development

Since the 1970s, economic development has been understood to be a process with many dimensions.

> **Economic development** is defined as a process where increases in real *per capita* output and incomes are accompanied by improvements in standards of living of the population and reductions in poverty, increased access to goods and services that satisfy basic needs (including food, shelter, health care, education, sanitation and others), improved gender equality, increasing employment opportunities and reduction of unemployment, and reductions of serious inequalities in incomes and wealth.

[4] Charles P. Kindleberger (1965) *Economic Development*, McGraw-Hill.

[5] David Morawetz (1977) *Twenty-Five Years of Economic Development: 1950–1975*, World Bank.

Human development

Thinking about development has progressed further, building on an even broader interpretation of development provided by Denis Goulet as early as 1971. Goulet defined three core values of development:[6]

- **Life sustenance** refers to access to basic services (merit goods) such as education and health care services, as well as satisfaction of basic needs like food, clothing and shelter.
- **Self-esteem** involves the feeling of self-respect; development provides individuals with dignity, honour and independence. Self-esteem is related to the absence of exploitation and dominance associated with poverty and dependence.
- **Freedom** involves freedom from want, ignorance and squalor; it is freedom to make choices that are not available to people who are subjected to conditions of poverty.

The economist Amartya Sen, who won the Nobel Prize in Economics in 1998 for his work on poverty and economic development, expands on Goulet's ideas, and sees improvements in human well-being as arising from a process of expanding freedoms:

> *'Development can be seen . . . as a process of expanding the real freedoms that people enjoy. Focusing on human freedom contrasts with narrower views of development, such as identifying development with the growth of gross national product, or with the rise in personal incomes, or with industrialisation, or with technological advance, or with social modernisation. Growth of GNI or of individual incomes can, of course, be very important as* means *to expanding the freedoms enjoyed by the members of the society. But freedoms depend also on other determinants, such as social and economic arrangements (for example, facilities for education and health care), as well as political and civil rights (for example, the liberty to participate in public discussion and scrutiny) . . .'*[7]

Sen's approach has been crystallised in the concept of *human development*, introduced in the first Human Development Report of the United Nations Development Programme (UNDP) in 1990.[8]

> *Human development* is a process of expanding human freedoms: the freedom to satisfy hunger; to be adequately fed; to be free of preventable illnesses; to have adequate clothing and shelter; to have access to clean water and sanitation; to be able to read, write and receive an appropriate education; to be knowledgeable; to be able to find work; to enjoy legal protection; to participate in social and political life; and, in general, to have the freedom to develop one's potential and lead a full and productive life.

Using the concept of human development, the UNDP makes a distinction between *income poverty* and *human poverty*. Income poverty occurs when income falls below a nationally or internationally determined poverty line. Human poverty involves deprivations and the lack of opportunities that allow individuals 'to lead a long, healthy, creative life and to enjoy a decent standard of living, freedom, dignity, self-esteem and the respect of others'.[9]

To understand the distinction between the two, consider a villager whose income increases, so that now he or she is able to purchase more goods and services. If there are no schools or health care services in the area, or if the village is infested with malaria, the higher income will be of little use in securing a higher standard of living. Income poverty is reduced, but human poverty cannot be lowered without measures to provide a broad range of social services to the entire population. On the other hand, if people on low incomes have access to education, health services, improved sanitation, improved water supplies, and so on, human poverty can be reduced even while income poverty remains.

[6] D. Goulet (1971) *The Cruel Choice: A New Concept on the Theory of Development*, Atheneum.

[7] Amartya Sen (2001) *Development as Freedom*, Oxford University Press. Emphasis in the original.

[8] The United Nations Development Programme (UNDP) is an agency of the United Nations designed to promote development in economically less developed countries and reduce poverty. The UNDP's Human Development Reports are compiled annually and provide statistical and other information on numerous issues relating to human development in less developed countries around the world.

[9] United Nations Development Programme, *Human Development Report 1997* (available at http://hdr.undp.org/en/reports/).

THEORY OF KNOWLEDGE 18.1

The values of economic development

Whereas the study of economic development is part of the social science of economics, it cannot be separated from normative values about what is believed to be good for development. As noted above, in 1971 Denis Goulet provided three core values of development, which have stood the test of time and formed the cornerstone of the meaning of economic and human development: life sustenance, self-esteem and freedom. On these three basic values, numerous others have been superimposed: poverty alleviation, improved income distribution, universal education, improved access to health care, improved employment opportunities, women's empowerment, institutional modernisation, and many more. All these represent means toward the ends of achieving the goals of life sustenance, self-esteem and freedom, and at the same time they are also ends in themselves (goals) because of the commonly held belief that they are good things for all societies in the world to achieve.

The meaning of development cannot be separated from these normative values. The Sustainable Development Goals (SDGs), all of which rest on normative values about what is good for societies, are goals of development, and at the same time they are also the way we measure how development advances, by use of the indicators that measure how much progress has been made toward achieving each goal.

How does economics as a social science reconcile itself with the presence of numerous values of development? The answer is that the scientific part of economics can be used to show what kinds of policies are most effective, or not effective in achieving a particular goal; or what policies are consistent with each other or conflict in the pursuit of one or more goals; or what is the least costly way to achieve some goal; or which members of a society are most likely to benefit from advancing toward a goal; and so on.

Thinking points

- Is 'development' an appropriate concept to reflect the ideas and values discussed above, or might another concept, like 'progress', or 'transformation', or something else be more appropriate?
- There are very many different cultures around the world, each of which has its own set of social customs and values. Examine each of the goals listed in the Sustainable Development Goals and try to determine whether there might be some values that would not be accepted by some cultures.
- Do you think there are any universal values for development, or values that are shared by all societies in the world?

Measuring economic development

Introduction to the use of indicators

Economic development, being complex and multidimensional, is not accurately reflected in any single measure. Economists therefore consider many different individual attributes or characteristics that distinguish countries according to their level of economic or human development.

Individual attributes and characteristics are measured by use of **indicators**. An indicator is a measurable variable that indicates the state or level of something being measured. For example, GDP *per capita* is an indicator of the level of output per person. The number of years of life expectancy is an indicator of a population's state of health. The proportion of a population that can read and write (literacy) is an indicator of the level of education. All these are attributes of economic or human development.

The data comprising economic or human development indicators are compiled by statistical services in every country and are made available to international organisations such as the World Bank and United Nations agencies. Indicators are extremely useful for:

- monitoring how a country changes (develops) over time with respect to the attribute measured by the indicator
- making comparisons between countries with respect to the attribute

- assessing how well a country is performing with respect to particular goals or targets of development (for example, an increase in the literacy rate indicates an improvement in educational level)
- devising appropriate policy measures to deal with specific problems.

In addition to individual attributes and their indicators, economists also use *composite indicators,* which are a summary measure of several dimensions or goals of development. We will consider both individual and composite indicators.

Single indicators

There are many hundreds of indicators used as measures of different characteristics of an economy and of dimensions of development. We will consider some important examples.

GDP *per capita* and GNI *per capita*

In Chapter 8, we learned that GDP is an indicator of the value of output produced within a country, and GNI is an indicator of the income (or value of output) received by the residents of a country, usually within a year. *Per capita* means that these values are calculated on a per person basis.

For most countries the difference in the sizes of GDP *per capita* and GNI *per capita* is not very large. This happens when inflows of income into a country are roughly balanced by income outflows. Otherwise, the difference can be significant. The factors of production that mainly account for differences are labour and capital.

When a country has many workers from other countries (labour) who send part of their wages back home (worker remittances), or foreign corporations (capital) that send their profits back home (profit repatriation), this decreases GNI relative to GDP so that domestic incomes received on average (GNI *per capita*) are lower than the value of output produced in the country (GDP *per capita*).

By contrast inflows of money into a country from workers abroad or from corporations located abroad increase the size of GNI relative to GDP, making domestic incomes on average (GNI *per capita*) higher than the value of output produced in the country (GDP *per capita*). It follows therefore that:

GNI *per capita* is a better indicator of the standards of living of a country, because it represents income per person received by the residents. GDP *per capita* is a better indicator of the level of output per person produced in a country.

Table 18.2 provides some examples of GNI *per capita* and GDP *per capita*. The last column, showing GNI as a percentage of GDP, is a convenient way to compare the two. If the value of the percentage is larger than 100, GNI > GDP; if it is smaller, GNI < GDP.

Two countries in the table that stand out are Ireland and the Philippines. Ireland's GNI is 77.5% of its GDP *per capita* because it has many multinational corporations, as well as foreign workers, who send profits and wage incomes back to their home countries. In the case of the Philippines, its GNI *per capita* is 122.5% of its GDP *per capita*, due to the important role played by worker remittances.

GDP *per capita* and GNI *per capita* in terms of PPPs

In Table 18.2, we compared GDP *per capita* with GNI per *capita* figures for individual countries. However, if we wanted to compare GDP *per capita* (or GNI *per capita*) *across countries*, the information in this table would lead to misleading conclusions because different countries have different *price levels*. This topic was discussed in Chapter 8 where we saw that that the same amount of money in a low-price country has greater purchasing power (can buy more things) than in a high-price country. This problem is resolved by use of purchasing power parity (PPP) exchange rates. PPPs eliminate the effect of price level differences, making GDP or GNI figures directly comparable across countries.

Table 18.3, showing GDP *per capita* calculated both by use of standard exchange rates (column 1) and by use of purchasing power parities (column 2), reveals an interesting pattern: for the poorer countries starting at the top of the table, GDP figures based on PPPs are higher than those based on exchange rates; for the wealthier countries at the bottom of the table, GDP figures based on PPPs are lower than those based on exchange rates.

	GNI *per capita* (US$)	GDP *per capita* (US$)	GNI as % of GDP*
High income countries			
United Kingdom	40 600	39 954	101.6
Japan	38 520	38 430	100.2
Switzerland	81 130	80 343	101.0
United States	59 160	59 928	98.7
Australia	51 360	53 793	95.5
Canada	42 790	44 871	95.4
Ireland	53 370	68 885	77.5
Middle and low-income countries			
Philippines	3660	2989	122.5
Lesotho	1210	1154	104.9
Pakistan	1580	1548	102.1
China	8690	8827	98.4
Bangladesh	1470	1516	97.0
Chad	640	662	96.7
Colombia	5980	6409	93.3
Indonesia	3540	3846	92.0
India	1790	1979	90.4
Kazakhstan	7970	9030	88.3
Russia	9220	10 749	85.8

* GNI as % of GDP is identical to GNI *per capita* as % of GDP *per capita*

Source: *World Bank, World Development Indicators https://data.worldbank.org/indicator*

Table 18.2: GNI *per capita* and GDP *per capita* in selected countries, 2017

	1 **GDP *per capita* (converted into US$ by use of exchange rates)**	2 **GDP *per capita* (converted into US$ by use of US$ PPP)**
Burundi	292	735
Pakistan	1548	5539
Philippines	2989	8360
China	8827	16 842
Argentina	14 398	20 829
Czech Republic	20 380	38 020
Japan	42 583	42 067
United States	59 928	59 928
Norway	75 704	62 183
Switzerland	80 343	66 307
World	10 749	17 100

Source: *World Bank, World Development Indicators*

Table 18.3: GDP *per capita* using exchange rates and purchasing power parities

The reason is that prices of goods and services on average tend to be lower in countries with low *per capita* GDPs, and higher in countries with high *per capita* GDPs. Consider two countries that produce an identical quantity of output, but that have different prices for this output. When the value of output is calculated in terms of US$ using exchange rates, it appears lower in the lower price country than in the higher price country, even though the quantity of output is the same.

This is exactly what happens in the real world. Column 2, using PPPs to convert GDP *per capita*, eliminates the impact on GDP of differing price levels, and as a result, the differences in *per capita* GDP between countries shrink enormously. Comparing Switzerland with Burundi, we see that Switzerland's GDP *per capita* based on exchange rates is 275 times greater than Burundi's; based on purchasing power parities, it is 90 times greater. The second comparison is a much better indicator of the differences in output produced in Burundi and Switzerland.

In the case of the United States, the two figures are identical, since it is the purchasing power of the US$ within the United States that is used as the basis for the PPP conversions.

Everything that has been said here about comparisons of GDP *per capita* across countries applies equally to GNI *per capita* (or any other output or income measure).

> Comparisons of GDP *per capita* (or GNI *per capita*) across countries require measures of *per capita* output or income based on conversions of national currencies into US$ by use of purchasing power parities (PPPs), to eliminate the influence of price differences on the value of output or income.

Purchasing power parity exchange rates are computed and published on a regular basis by several international bodies, including the Organisation for Economic Co-operation and Development (OECD), the European Union, the World Bank and United Nations agencies.

Health indicators

Three common health indicators are life expectancy at birth, infant mortality and maternal mortality. Data for all three, together with GNI *per capita* (US$ PPP), for selected countries are provided in Table 18.4. (We are using GNI *per capita* as this is a better indicator of living standards, and we are using values in US$ PPP to eliminate the influence of price-level differences across countries.)

Life expectancy at birth is the average number of years of life in a population. It is one of the most commonly used indicators of development. *Infant mortality* refers to the number of infant deaths from the time of birth until the age of one, per 1000 live births. *Maternal mortality* refers to the number of women who die per year as a result of pregnancy-related causes, per 100 000 live births.

Table 18.4 shows that higher levels of GNI *per capita* (US$ PPP) tend to be linked with higher life expectancies, and lower infant and maternal mortalities. This is what we would expect, since higher income countries have more resources to provide the necessary services and appropriate living conditions for their populations. However, there are very wide departures from this broad pattern, suggesting that *income per capita is not the only factor that determines health outcomes in a country.*

Among more developed countries, the United States stands out for its lower life expectancy and higher infant and maternal mortalities compared to other more developed countries.

Among less developed countries, we find some very surprising health outcomes. Burundi with a GNI *per capita* less than half that of Sierra Leone has infant mortality and maternal mortality rates that are almost half that of the latter. Moldova with GNI *per capita* lower than those of Angola and India has an infant mortality rate, but especially maternal mortality rate, far lower than the other two countries. The table has more such examples.

How is it possible that some countries have managed to achieve far better health outcomes than others with similar or even lower incomes *per capita*? The answer is that for any given level of income *per capita*, life expectancy is *higher*, and infant mortality and maternal mortality are *lower*, when there are:

- adequate public health services (such as immunisation, provision of health information and education), and prevention of communicable diseases (such as malaria, tuberculosis, HIV/AIDS)
- adequate health care services with broad access by the entire population
- a healthy environment, including safe drinking water, sewerage and sanitation, and low levels of pollution
- an adequate diet and avoidance of malnutrition
- a high level of education of the entire population
- absence of serious income inequalities and poverty.

Country	GNI per capita US$ PPP 2017	Life expectancy at birth (years) 2017	Infant mortality rate (per 1000 live births) 2017	Maternal mortality ratio (per 100 000 live births) 2017
Norway	64 760	83	2	5
United States	61 120	79	6	14
Finland	46 880	81	2	3
United Kingdom	44 090	81	4	9
Japan	43 540	84	2	5
Greece	28 640	81	4	3
Turkey	27 640	76	10	16
Russia	25 120	72	7	25
Chile	23 780	80	6	22
China	16 800	76	8	27
Sri Lanka	12 520	75	8	30
Armenia	10 060	75	11	25
India	6950	69	32	174
Angola	6450	62	54	477
Moldova	6100	72	13	23
Zambia	3900	62	42	224
Chad	1920	53	73	856
Uganda	1820	62	35	343
Sierra Leone	1510	52	82	1360
Burundi	730	58	43	712
World	17 043	72	29	216

Source: *World Bank, World Development Indicators*

Table 18.4: Health indicators in selected countries in relation to GNI *per capita* in US$ PPP

Therefore, health outcomes depend a lot on how well countries achieve these objectives. For example:

- Health outcomes in the United States may be due to inequalities in income and education resulting in pockets of poverty, connected to poor housing and living conditions, poor nutrition and health, and insufficient access to medical care (due to lack of medical coverage). Such factors result in worse health outcomes among low-income groups, which lowers the average over the entire American population.
- Health outcomes in countries like Burundi and Moldova (and many others) are due to government policies placing a high priority on public health and the provision of health care services for low-income groups, as well as on education.
- Health outcomes in sub-Saharan African countries are due to a very large extent to the disastrous impacts of HIV/AIDS, as well as problems with sanitation, safe drinking water, lack of education and information, poor public health and health care services, and premature deaths due to diseases that are both preventable and treatable.

The discussion of health indicators illustrates that:

- GNI *per capita* (or any other income or output measure) is an insufficient indicator of health outcomes.
- Limited resources, due to low GNI *per capita*, are not always the most important cause of poor health outcomes. Most (if not all) countries, both more and less developed, can do more with their available resources to meet economic development goals. They can reallocate resources towards provision of more social services and merit goods, improving the institutions through which these services are delivered, as well as reducing poverty.
- Some development issues apply not only to developing, but to developed countries as well, because of the presence of poverty in wealthy societies that make people on low incomes subject to similar deprivations as poor people in developing economies.

Education indicators

Education indicators measure levels of educational attainment. There are many such indicators of which three are shown in Table 18.5.

The adult literacy rate measures the percentage of people aged 15 or more in the population who can read and write. Primary school enrolment measures the percentage of school-age children who are enrolled in primary school (elementary school). Lower secondary school enrolment measures the percentage of children enrolled in the lower years of secondary school (high school).

Table 18.5 shows that as income *per capita* increases, all three indicators tend to increase. However, as in the case of health indicators, there are many exceptions, involving countries with relatively low incomes that have high levels of educational attainment, especially in adult literacy and primary school enrolment.

There are two main reasons for these exceptions. One is that countries of the former Soviet Union and other former communist countries have very good education outcomes because historically, communist governments placed a high priority on education. This explains the high achievements of Armenia and Moldova relative to their GNI *per capita*. The second reason is that some governments have made a special effort to provide education services to their populations. We therefore see countries like Bolivia and Uganda with very good education outcomes compared to countries with comparable or higher incomes *per capita*. Perhaps most striking are the primary enrolment rates of Zambia, Uganda and Burundi, among the poorest countries in the world, as well as Burundi's literacy rates, achieved in spite of very low GNI *per capita*.

The case of secondary enrolment is different. Countries that are economically less developed must use their scarce resources to provide primary education to achieve universal literacy, which is an important precondition for economic growth and development. Secondary education is less of a priority for low-income countries. Therefore, it is not surprising that countries with very good achievements in primary enrolment lag behind in secondary enrolment.

Countries can achieve universal literacy and universal primary education even if they have relatively low *per capita* incomes, provided their governments allocate enough resources to education services, and ensure that all children have access to these.

Economic inequality indicators

Economic inequality indicators were presented in Chapter 12. They include Lorenz curves, Gini coefficients, poverty lines, minimum income standards, and the Multidimensional Poverty Index.

REAL WORLD FOCUS 18.1

Rich country lifestyle diseases threaten progress in Africa

Figure 18.1: An African baby girl being weighed on a scale

Life expectancy in Africa is longer than ever before. In sub-Saharan Africa it is 66 years for women and 62 years for men, or 64 years on average. In the year 2000, life expectancy was 53 years. This has been accompanied by a huge decrease in mortality of children younger than five, where the number of deaths fell from 45% in 1950 to 10% in 2017.

It is estimated that about one-third of the improvements have come from increasing *per capita* incomes, one-third from improvements in education, and one-third from changes that include technological improvements such as vaccines and improved disease control methods. Additional reasons may include improved job opportunities, better working conditions and safer housing.

Still, Africa has far to go to reach the levels of health of more developed countries where life expectancy on average is 82 for women and 76 for men. The gap between Africa and developed countries arises mainly from the prevalence of diseases that are curable or preventable or both, having their root in poverty (see Chapter 19).

Aside from these factors, a major study on the leading causes of death in sub-Saharan Africa found that increasingly, deaths are caused by lifestyle factors that are imported from rich countries. In rich countries, non-communicable diseases including heart attacks and stroke are among the major causes of death. These illnesses are strongly linked to lifestyle factors such as alcohol consumption, unhealthy diets and lack of exercise, that result in high blood pressure, diabetes and obesity.

For people of ages 50–69, the major causes of death in sub-Saharan Africa currently are stroke and heart attacks, which are associated with high blood pressure, diabetes and obesity. This suggests that lifestyle factors may start to erode the gains made in life expectancy over the last decades.

Source: *People are living longer in Africa but the rise of lifestyle diseases threatens progress https://theconversation.com*

Applying your skills

Consider and discuss the possible policies that governments and international aid organisations can adopt in order to address the problems caused by lifestyle factors affecting life expectancy in Africa.

Country	GNI *per capita* US$ PPP 2017	Total adult literacy rate (% of people aged 15 and above) 2015–2016	Primary school enrolment (% of children of official school age) 2010–2016	Lower secondary school enrolment (% of children of official school age) 2010–2016
China	16 800	96.4	97	–
Colombia	14 120	94.7	94	77
Peru	12 900	94.2	92	86
Sri Lanka	12 520	91.9	–	–
Ecuador	11 350	94.4	97	–
Armenia	10 060	99.7	100	98
Morocco	8050	68.5	89	–
Bolivia	7350	92.5	97	–
India	6950	71.2	83	–
Angola	6450	71.1	76	31
Moldova	6100	99.4	–	–
Zambia	3900	63.4	87	49
Chad	1920	22.3	50	13
Uganda	1820	78.4	87	17
Sierra Leone	1510	48.1	76	36
Burundi	730	85.6	85	11

Source: *GNI per capita US$ PPP World Bank, World Development Indicators*
Literacy rates CIA The World Factbook
Primary school enrolment Unicef, Data and Analytics Section, Division of Data Research and Policy

Table 18.5: Education indicators in relation to GNI *per capita*

Social inequality indicators

There are very many indicators of social inequality. The list below presents just a few examples of these:

- Adolescent fertility rates
- Prevalence of undernourishment
- Inequality in life expectancy
- Inequality in education
- Gender inequalities
- Populations vulnerable to poverty
- Child malnutrition
- Infants lacking immunisation
- Child labour
- Old-age pension recipients
- Homeless people due to natural disaster
- Birth registration

REAL WORLD FOCUS 18.2

Child labour

Figure 18.2: Istahua, El Salvador. A ten-year-old boy works at a brick factory

Goal 8 of the Sustainable Development Goals 'Decent Work and Economic Growth' and its associated Target 8.7 asks the international community to

> *'Take immediate and effective measures to eradicate forced labour, end modern slavery and human trafficking and secure the prohibition and elimination of the worst forms of child labour, including recruitment and use of child soldiers, and by 2025 end child labour in all its forms.'*

At Lake Volta in Ghana, one in six children between the ages of 6 and 14 do hazardous work 17 hours a day, facing starvation and beatings as punishment. This results from poverty, with work offering money to the family and the promise of learning a trade, but perpetuates poverty by depriving children of an education and causing psychological damage.

The International Labour Organization of the United Nations began monitoring child labour in 2000. Since that time major progress was made in reducing child labour, resulting in a reduction of child labour by 94 million children in the period 2000–2016. However the progress has stalled in the last few years. It is estimated that 152 million children, including 64 million girls and 88 million boys are engaged in child labour, with nearly half of them, 73 million, involved in work that endangers their health, safety or moral development.

Note that child labour does not include legal forms of child employment, which involves an additional 66 million children.

Children aged 5 to 11 are the largest share of child labour and also the largest share of dangerous work. Many of these children are completely deprived of education. However, even when the children do attend school, research indicates that their activities as labourers interferes with their studies and they perform poorly at school.

48% 5–11 years old 28% 12–14 years old 24% 15–17 years old	58% boys 42% girls	70.9% agriculture 11.9% industry 17.2% services	43.0% low income countries 38.4% lower middle income countries 17.3% upper middle income countries 1.3 % high income countries

Table 18.6: 152 million children in child labour

CONTINUED

Policies in the fight against child labour include the following:

- Provision of free and compulsory education up to a minimum age.
- Provision of social protection systems to deal with poverty (for example cash transfers, unemployment protection, old-age pensions, (see Chapter 20) in order to reduce dependence on children to supplement family income.
- Labour market policies to tackle youth unemployment, including vocational and technical training, with an emphasis on workers' rights and protection from hazardous work conditions.

Source: *Children in employment www.ilo.org; www.borgenmagazine.com*

Applying your skills

Use a poverty cycle diagram to explain how child labour traps children in a lifetime of poverty.

Energy indicators

Energy indicators are similarly very numerous. Examples include the following:

- renewable energy consumption
- access to electricity
- electric power consumption.

A joint publication of five organisations, *Indicators for Sustainable Development: Guidelines and Methodologies*,[10] has identified 30 indicators, reduced from an original set of more than 130, to be used to facilitate national policy-making and measurement of performance. These indicators are classified according to three dimensions: social, economic and environmental. For example:

Social dimension

- Share of households (or population) without electricity or commercial energy, or heavily dependent on non-commercial energy.
- Share of household income spent on fuel and electricity.

Economic dimension

- Energy use *per capita*.
- Renewable energy share in energy and electricity.

Environmental dimension

- Air pollutant emissions from energy systems.
- Rate of deforestation attributed to energy use.

Environmental indicators

Environmental indicators help provide a description of developments affecting the environment that can be used to monitor changes and progress toward meeting environmental objectives. Such indicators are also very numerous, having become increasingly popular since the 1990s. Some examples include the following:

- CO_2 emissions per unit of GDP or *per capita*
- Emissions of other hazardous substances
- Bird species threatened
- Fish species threatened
- Measures of ozone layer depletion
- Measures of waste generation
- Measures of waste water treatment
- Measures of intensity of water use

[10] International Atomic Agency, United Nations Department of Economic and Social Affairs, International Energy Agency, Eurostat, European Environment Agency https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1222_web.pdf

TEST YOUR UNDERSTANDING 18.2

1. a. Distinguish between economic growth and economic development.
   b. Outline the dimensions of economic development.
2. a. Explain the difference between GDP *per capita* and GNI *per capita*.
   b. Outline the factors that account for differences between the two measures.
   c. Explain why Ireland's very high GDP *per capita* gives a misleading impression of average standards of living in Ireland.
   d. Examine the figures for Russia in Table 18.2 and try to determine what might account for the difference between its GDP *per capita* and GNI *per capita*.
3. Outline the meaning of purchasing power parities (PPPs) noting why they are important for making valid comparisons of GDP or GNI across countries.
4. a. Using Table 18.4, explain some differences in health outcomes between more and less developed countries.
   b. Provide examples of countries that have achieved good health outcomes relative to their level of income.
   c. Provide examples of countries that could likely achieve better outcomes, even with their given level of income.
   d. Use a *PPC* diagram to illustrate your answers.
5. Answer all the parts of question 4 using Table 18.5, referring to education outcomes.
6. Identify examples of
   a. economic inequality indicators,
   b. social inequality indicators,
   c. energy indicators, and
   d. environmental indicators.

Composite indicators

Being multi-dimensional, economic development cannot be adequately measured by any single indicator. This problem led two famous economists – Mahbub ul Haq, a highly influential Pakistani development economist, and Amartya Sen, an Indian economist who won the Nobel Prize for his work in economic development – to develop **composite indicators**, which are summary measures of more than one dimension of development. By including more than one dimension, composite indicators are more accurate measures of development. Their work was carried out at the United Nations Development Programme (UNDP), which since 1990 produces a Human Development Report every year with analyses of various development issues as well as statistical information, including information on the Sustainable Development Goals, indicator targets and composite indicators. We will study three of the composite indicators prepared by the UNDP: the Human Development Index (HDI), the Inequality-adjusted Human Development Index (IHDI), and the Gender Inequality Index (GII).

In addition we will examine the Happy Planet Index, introduced in Chapter 8, which is prepared by New Economics Foundation.

Composite indicators are usually expressed as an 'index' or a set of numbers showing the relative position of a variable in a list. This will become clearer in the discussion below.

The Human Development Index

The **Human Development Index (HDI)**, the best-known and most widely used index of the UNDP, is based on the concept of *human development* discussed earlier in this chapter. The HDI measures average achievement in three dimensions: a long and healthy life, access to knowledge and a decent standard of living. As of 2010, these three dimensions are measured by the following indicators:

- a long and healthy life is measured by life expectancy at birth
- access to knowledge is measured by mean years of schooling and expected years of schooling
- a decent standard of living is measured by GNI *per capita* (US$ PPP).

Each dimension is expressed as a value between 0 and 1, with 0 being the lowest possible value for the dimension, and 1 being the highest.

The composite index is the average over the three dimensions. Each country receives an HDI value from

Country	HDI rank 2017	Human Development Index 2017	Life expectancy at birth	Expected years of schooling	Mean years of schooling	GNI *per capita* US$ PPP
Spain	25	0.891	83.3	17.9	9.8	34 258
Luxembourg	26	0.914	82.0	14.0	12.1	65 016
Tajikistan	127	0.650	71.2	11.2	10.4	3317
Namibia	128	0.647	64.9	12.3	6.8	9387
India	129	0.640	68.8	12.3	6.4	6353
Myanmar	147	0.578	66.7	10.0	4.9	5567
Nepal	148	0.574	70.6	12.2	4.9	2471

Source: United Nations Development Programme, Human Development Report 2018

Table 18.7: Human Development Index (GDI) and GNI per capita (US$ PPP) for selected countries

0 to 1, and the countries are ranked according to their HDI values.

HDI ranks and HDI values for selected countries, together with their corresponding GNI *per capita* (in US$ PPP), appear in Table 18.7, where selected countries are listed in order of declining HDIs. Countries have been selected to show how it is possible to achieve similar levels of human development with very different levels of GNI *per capita*.

For example, Myanmar and Nepal have similar HDIs, indicating that they have attained approximately the same level of human development, yet Nepal has accomplished this with a GNI per capital that is less than half of Myanmar's GNI per capita. The same can be said for Tajikistan, Namibia and India.

Comparisons between HDIs and GNI *per capita* confirm the points made earlier:

- GNI (or GDP) *per capita* used alone can be a poor measure of the different dimensions of development.
- Many countries, even with their given levels of GNI *per capita*, are capable of making significant improvements in the well-being of their populations by making different choices regarding the resources allocated to health, education and other services or merit goods.
- Economic and human development issues apply not only to developing countries, but to developed countries as well.

The HDI is very useful as a tool for governments wishing to devise policies focusing on economic and human development, and *it is far superior to single indicators as a measure of development*. However, the HDI, too, has its shortcomings. This is because economic and human development are much broader concepts with more dimensions than are reflected in the HDI. The HDI does not provide us with information about income distribution, malnutrition, demographic trends, unemployment, gender and other inequalities, political participation, etc.

Inequality-adjusted Human Development Index

The **Inequality-adjusted Human Development Index (IHDI)** measures human development in the same three dimensions as the HDI, but each dimension is adjusted for *inequality* in the corresponding dimension. The IHDI attempts to measure losses in human development that arise from inequality. If there were perfect equality in income, health and education, the IHDI would be exactly equal to the HDI. Where there are inequalities, the IHDI is lower than the HDI, and the greater the inequalities, the lower the IHDI relation to the HDI.

Interestingly, the IHDI is lower than the HDI in the case of all countries examined by the United Nations Development Programme (UNDP). Table 18.8 shows the five countries whose HDI has decreased the least as well as the five countries whose HDI decreased the most on account of inequalities. The column showing 'Overall loss' shows the percentage of the HDI that was

Country	HDI	IHDI	Overall loss (%)	Gini coefficient
Japan	0.909	0.876	3.6	32.1
Czech Republic	0.888	0.840	5.3	25.9
Finland	0.920	0.868	5.6	27.1
Slovenia	0.896	0.846	5.6	25.4
Iceland	0.935	0.878	6.0	25.6
Gambia	0.460	0.289	37.2	35.9
Chad	0.404	0.249	38.3	43.3
Haiti	0.498	0.304	39.0	41.1
Central African Republic	0.367	0.212	42.1	56.2
Comoros	0.503	0.275	45.3	45.3
World	0.728	0.582	20.0	-

Source: United Nations Development Programme, Human Development Report 2018

Table 18.8: Inequality-adjusted Human Development Index (IHDI) in relation the Human Development Index (HDI) and Gini coefficients

lost due to the presence of inequalities. The last column shows the Gini coefficients for each of the countries. The data suggest two patterns. One is that the countries that lost the least of their HDI values are countries with a high human development while those that lost the most have a relatively low human development. The other is that countries that lost the least have a relatively more equal distribution of income as indicated by their Gini coefficient. This is what we would expect since income inequality is one of the dimensions of the IHDI. According to the UNDP, 'At the global level, inequality in income contributes the most to overall inequality, followed by education and life expectancy.'[11]

It may be noted that the IHDI is an improvement over the HDI in that it includes the dimension of inequality. However it, too, is an incomplete measure of the level of economic development for the same reasons as the HDI noted earlier.

Gender Inequality Index

The **Gender Inequality Index (GII)** measures inequalities between the genders in three dimensions measured by the following indicators:

- reproductive health is measured by
    - the maternal mortality ratio (death per 100 000 live births; see Table 18.4)
    - the adolescent birth rate (births per 1000 women ages 15–19)
- empowerment is measured by
    - the share of parliamentary seats held by women
    - the proportion of women in the total population with at least some secondary education
- labour market participation is measured by the proportion of women in the labour force.

The GII measures the loss in human development of women due to inequalities in these areas. The higher the GII, the greater the gender inequality. Table 18.9 shows values for the GII according to groups of countries.

[11] United Nations Development Programme, Human Development Report 2018, p 4
http://hdr.undp.org/sites/default/files/2018_human_ development_statistical_update.pdf

Country group	Gender Inequality Index
Sub-Saharan Africa	0.569
Arab States	0.531
South Asia	0.515
Latin America and the Caribbean	0.386
East Asia and the Pacific	0.312
Europe and Central Asia	0.270
OECD countries	0.186
World	0.441

Source: *United Nations Development Programme, Human Development Report 2018*

Table 18.9: The Gender Inequality Index according to groups of countries

Happy Planet Index

The **Happy Planet Index (HPI)** was discussed in Chapter 8. It is suggested that you reread this section. As you may recall, the architects of the HPI argue that GDP (or similarly GNI) is not a good indicator of well-being. By contrast, the HPI is proposed as a measure of sustainable well-being, that takes into consideration life expectancy, people's feelings about their own personal well-being, with adjustments for inequalities and ecological footprint. The HPI ranks countries on the basis of their performance based on these indicators. Table 18.10 shows the HPI ranks of selected countries, and compares them with GNI per capita as well as HDI ranks.

As the table shows, countries with high ranks on the basis of the GNI per capita or HDI ranks do not rank so highly on the HPI, in fact there is little correspondence between the three indicators shown in the table. For the most part this is due to the ecological footprints of high-income countries such as Germany, Canada, Australia and Luxembourg, which work to substantially lower their rank in the HPI. You may note too that Germany with a GNI *per capita* that is more than seven times greater that of India, has almost the same HPI rank as India. This is also due to Germany's strong ecological footprint compared to India's very good performance in this respect.

The Happy Planet Index provides information that is of a very different nature to the UNDP indices, as it is the only one that adjusts for unsustainable resource use through its inclusion of the ecological footprint. Yet it too is incomplete as it does not account for other important dimensions of development, while the concepts of well-being and ecological footprint remain controversial.

Country	Happy Planet Index Rank	Happy Planet Index	Human Development Index Rank	GNI per capita US$ PPP
Costa Rica	1	44.7	63	14 636
Indonesia	16	35.7	115	10 846
Germany	49	29.8	4	46 136
India	50	29.2	129	6 353
Canada	85	23.0	12	43 433
Malawi or Tanzania	98	22.1	170	1064
Australia	105	21.2	3	43 560
Luxembourg	139	13.2	26	65 016

Source: *Happy Planet Index http://happyplanetindex.org*
Human Development index and GNI per capita United Nations Development Programme, Human Development Report 2018 http://hdr.undp.org/sites/default/files/2018_human_development_statistical_update.pdf

Table 18.10: Happy Planet Index in relation to GNI *per capita* and HDI ranks for selected countries

TEST YOUR UNDERSTANDING 18.3

1 Distinguish between single and composite indicators. Provide some examples of each.

2 a Outline the three dimensions of development measured by the Human Development Index (HDI).

b Explain the advantages of using the Human Development Index over GDP or GNI *per capita* as a measure of economic and human development.

3 a Identify the meaning of a high HDI rank and HDI value in terms of the level of human development.

b Explain what it means if a country's GNI *per capita* rank is (i) higher than its HDI rank, and (ii) lower than its HDI rank.

4 a Explain the difference between the HDI and the IHDI.

b Explain why the IHDI is lower than the HDI for all countries.

c Explain the relationship between the size of the Gini coefficient appearing in Table 18.8 and the extent of the loss of the HDI for the countries appearing in the table.

5 a Outline the dimensions of the Gender Inequality Index (GII) and the indicators used to measure these.

b Identify the meaning of a high GII value in terms of the extent of gender inequality.

6 a Describe the dimensions of the Happy Planet Index.

b Explain the main factor that accounts for the relatively low ranks of economically more developed countries and relatively high ranks of some economically less developed countries.

7 Use examples of individual or composite indicators (you may use the tables in the text) to illustrate the following points:

a GNI *per capita* is sometimes a poor indicator of levels of economic and human development.

b Many countries around the world can do more to promote the well-being of their populations through a reallocation of resources, even in the absence of economic growth.

c Some economic development issues apply to more developed countries as well as less developed countries.

d Countries may be more developed with respect to some indicators and less developed with respect to other indicators.

Strengths and limitations of measures of economic development

Individual and composite indicators, used alone or in combination with *per capita* GNI (or GDP) statistics, are enormously useful as measures of different aspects of development.

However, as economic development is a complex and multidimensional process, it cannot be accurately represented by any one measure. Each single indicator, or group of indicators in a composite index, can only capture that part or parts of development that it measures. Therefore, we cannot rely on any one measure to obtain a complete picture of the level of development of a country. Very often it is necessary to combine the use of many indicators to obtain an overall picture.

Moreover, since different single or composite indicators measure different aspects of development, they sometimes present conflicting perspectives. For example, the Human Development Index (HDI) conflicts in part with the Inequality-adjusted Human Development Index (IHDI), and both of these conflict with the Happy Planet Index (HPI). The conflicting perspectives result in different values and rankings of the countries, depending on their performance in the particular dimensions of development that the composite indicator is measuring.

In addition, both single and composite indicators are subject to limitations arising from the fact that they are based on statistical information, which poses a distinct set of problems:

- Some countries have a limited capacity for collection of statistical data.
- Data are not fully available in many countries.
- In some situations (though not always) where data are missing international agencies try to come up with estimates but these may not be accurate.
- Very often international agencies do not have access to all the most recently available data with the result that data of different years may be compared against each other.
- Definitions of variables and methods used by statistical services vary from country to country, despite efforts by international organisations (such as the World Bank and United Nations agencies) to achieve standardisation. This means that the data may not always be directly comparable from country to country.

These statistical problems mean that the indicators cannot always be precise and should be used as rough guides of trends over time or differences between countries, rather than as very precise measures.

Relationship between economic growth and economic development

Economic growth can occur without economic development. It was this experience for many countries in the 1950s and 1960s that resulted in rethinking and redefining economic development. Can economic development occur without economic growth?

Some economic development is possible in the absence of rapid growth, if appropriate policies are followed to provide access to basic social services for the poor. The production possibilities model we studied in Chapter 1 shows how this can occur. In Figure 18.3, an economy produces some combination of industrial goods (measured on the vertical axis) and merit goods (measured on the horizontal axis). The merit goods include education, health care services, sanitation, and clean water supplies, made available to people on low incomes, who would not otherwise have access to them. (For simplicity it is assumed that actual output is at some point on the *PPC*). An economy that does

A → B: no economic growth with some development
B → C: economic growth with no development
B → D or E: economic growth with development

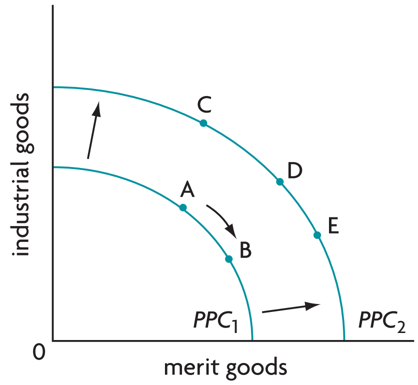


Figure 18.3: Economic growth and economic development

not experience growth can still achieve some economic development, by reallocating its resources, cutting back on industrial production and increasing merit goods production; this would entail a movement along PPC_1 from point A to a point like B.

Over long periods of time, the possibilities for improving the population's well-being by moving along the same *PPC* will be exhausted, and further improvements will depend on outward *PPC* shifts, such as from PPC_1 to PPC_2 in Figure 18.3. Such outward shifts, representing economic growth as an increase in production possibilities, are therefore necessary for economic development to be maintained. Growing output *per capita* translates into higher incomes and an improved ability to provide the goods and services needed by the population. However, economic growth *does not guarantee that economic development will occur*. If an economy moves from point A to point C, for example, there is little if any increase in merit good provision.

In this chapter we have seen several examples of countries that have greater achievements in education and health than other countries with higher levels of real GNI *per capita*. We have also seen examples of countries that succeeded in reaching a Human Development (HDI) value and rank similar to those of other countries with substantially higher GNIs per capita. These examples illustrate that it is possible to achieve some economic development by allocating resources to activities that improve living standards for the broader population.

TEST YOUR UNDERSTANDING 18.4

1 a Explain why it is easier to measure economic growth than economic development.

b Identify some difficulties involved in measuring economic development.

2 Use the *PPC* model to

a show how a country can experience growth without economic development,

b show how it can achieve some economic development without economic growth, and

c explain why economic development over long periods of time requires economic growth.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives this chapter.

1 Choose a Sustainable Development Goal that you are interested in and research its targets and indicators. Explain how the targets are related to the goal, and how the indicators can be used in order to measure a country's performance with respect to achieving the targets and the goal.

2 Select two countries of your choice with a comparable level of GNI per capita (US$ PPP). For each of these conduct an investigation of its

a Human Development Index (HDI),

b Inequality-adjusted Human Development Index (IGDI),

c Gender Inequality Index (GII), and

d Happy Planet Index.

Examine how the two countries compare with respect to their ranks in each of the indices. Then investigate the dimensions of each of the indices for your two countries and compare your results. You should be able to arrive at important conclusions regarding what dimensions of each index are responsible for any differences you observe in the index ranks of your two countries. For sources of information you may use the sources appearing in the tables of this chapter, or any more recent version of these publications.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 19

Barriers to economic growth and economic development

BEFORE YOU START

- What factors do you think prevent certain groups of people from climbing out of poverty?
- What factors do you think keep poor countries relatively poor?

Developing countries face numerous barriers to economic growth and economic development. Some of these are due to domestic factors while others are rooted in relations with other countries and the international economy. Whatever the case, it is important to identify what these are in order to best identify policies to overcome them.

Photo: Yongoro, Sierra Leone. Woman walking on the main road of Yongoro, near Freetown

19.1 Poverty cycles (or traps)

LEARNING OBJECTIVES

After studying this section you will be able to:

- explain the meaning of poverty cycles, also known as poverty traps (AO2)
- draw a diagram showing a poverty trap where poverty is perpetuated (AO4)

We learned about poverty in Chapter 12. While all countries in the world have poverty, most of extreme poverty (defined as living on less than $1.90 per day) is concentrated in developing countries. Poverty is caused by many factors, but one of the causes of poverty in some situations can be poverty itself. When conditions of poverty feed on themselves and create more poverty, they give rise to the poverty cycle, or trap.

Understanding the poverty cycle

People who are very poor spend their entire incomes just on essentials (food and other essential items), and often even this is not enough for survival. The physical capital they have is very low, whether this is farm tools, roads, water supplies or sanitation systems. Their human capital is very low: they have little if any education, low levels of skills and poor levels of health. Their natural capital often becomes depleted as they destroy their natural environment in an effort to survive (depletion of minerals in the soils, cutting of forests, overfishing of lakes, rivers and oceans, etc.). To come out of their state of poverty, they need more capital: physical, human and natural. The new capital can only be created by saving, which would enable them to make the necessary investments to increase their capital, but since all their income is spent on necessities, there is nothing left over to save; therefore, they cannot make investments in the capital they need. As a result, they are trapped in a situation where their poverty leads to more poverty, in a cycle. The poverty cycle is illustrated in Figure 19.1.

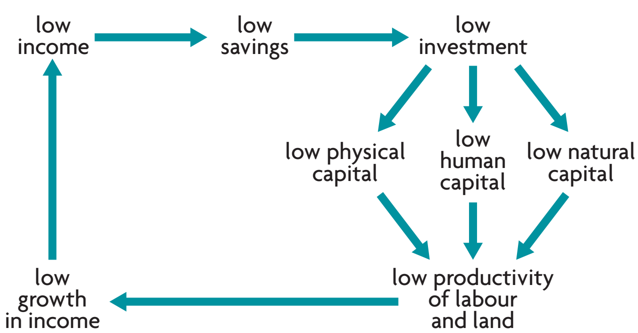


Figure 19.1: The poverty cycle (poverty trap)

A **poverty cycle (poverty trap)** arises when low incomes result in low (or zero) savings, permitting only low (or zero) investments in physical, human and natural capital, and therefore low productivity of labour and of land. This gives rise to low, if any, growth in income (sometimes growth may be negative), and hence low incomes once again. A poverty cycle may occur in a family, a community, a part of an economy, or in an economy as a whole. An important feature of the poverty cycle is that poverty is transmitted from generation to generation.

How poverty is transmitted across generations

There are a number of ways that poverty is transmitted across generations:

- People often cannot afford to send their children to school, either because the children work to supplement the family income, or because the parents cannot afford transport costs to school or the school fees.
- They cannot afford the necessary medical care for themselves or for their children, and sometimes cannot provide enough food for the family, leading to malnourished and physically disadvantaged children.
- They often have large families, whether because they see children as a source of additional income (if the children work), or as a source of security in old age, or because they do not have access to family planning services. Large families increase the level of poverty, as the income of the parents must be stretched to cover the needs of more people.

In all these cases, the children are penalised for life, as they grow into adulthood lacking skills, often unable to realise their full health potential, and condemned to low productivity and low incomes.

REAL WORLD FOCUS 19.1

The vicious cycle of poverty and population growth: the demographic trap

Evidence indicates that having many children increases poverty. Yet it is also the case that poor people tend to have more children. The reasons are many and varied: children may be viewed as a source of unpaid labour (on the farm), income (if the child is sent out to work), old-age security of the parents (the children, males in particular, are expected to take care of the parents in their old age); in addition, the children may be viewed as a source of prestige for the parents; or they may result from lack of access to contraception and family planning services.

Figure 19.2: Amazon rainforest, Brazil. Amazon riverine family - *ribeirinho* - living near the rivers with fishing as the main source of living, also cultivating small clearings for their own consumption

There is, therefore, a close and mutually reinforcing relationship between poverty and population growth: poor people have more children, and more children keep poor families trapped in poverty; this is known as the *demographic trap*. It is closely related to the poverty cycle or trap.

Experiences of many countries around the world show that there is a clear relationship between birth rates and economic growth and development: as countries grow and develop, birth rates fall. This relationship has given rise to the simplistic expression: 'Development is the best contraception.' It is simplistic because it begs the question: what aspects of growth and development lead to lower birth rates? Is it rising incomes, more education, increased contraception, increased employment opportunities, or another?

While there are many factors contributing to lower birth rates, the most important ones involve education and job opportunities for women. These are even more important than higher income *per capita* in order to break out of the demographic trap. When girls receive an education that prepares them for the job market, and when they join the labour force, they are far more likely to have fewer children. Reasons include later marriage, greater reproductive choice, increased awareness of the economic benefits of having fewer children (i.e. the ability to provide more per child if there are fewer children), and reduced need to view children as sources of income and old-age security. Nobel Prize-winning development economist Amartya Sen notes that in some of the richest districts of India, where women do not work and tend to receive less education, birth rates are much higher than in some far poorer districts where women have much higher literacy rates and greater labour force participation.[1]

Applying your skills

1. Outline why the demographic trap is a barrier to growth and development.
2. Explain the role of education and job opportunities for women in breaking out of the demographic trap.
3. Explain how the demographic trap is related to the poverty cycle (after reading about the poverty cycle).

[1] Amartya Sen, (1999) *Development as Freedom*, Oxford University Press.

Moreover, poor people who are unable to buy or invest in modern agricultural inputs (fertilisers, irrigation facilities, improved seeds) because their incomes are too low are forced to overuse their land, thus depleting the soil of essential nutrients, with the result that their children will be forced to work on soils of poorer quality that have lower yields (lower output per unit of land).

Since poor people cannot make investments because they do not have enough savings, it would help if they could borrow. However, banks do not usually lend to poor people, who lack the necessary collateral (see Section 19.3 below). Therefore, the poor remain without access to the credit that could help raise them out of their poverty, and poverty is carried into the next generation.

Breaking out of the poverty cycle

Poor people and poor communities trapped in a poverty cycle cannot emerge from this on their own. They require the intervention of the government, which must undertake investments in human capital (health services, education, nutrition), physical capital in the form of infrastructure (sanitation, water supplies, roads, power supplies and irrigation), and natural capital (conservation and regulation of the environment to preserve environmental quality). Further, the government must take the necessary steps to ensure that poor people can participate in private sector activities, such as ensuring access to credit so that the poor can borrow to finance private investments. However, the public investments needed to break out of the poverty trap depend on the availability of government revenues. What if the country is so poor and overall savings so low that the government does not have the revenues required to undertake the necessary investments? Then an entire country is trapped in a poverty cycle. There are several countries in sub-Saharan Africa that are trapped this way.

When an entire nation is trapped in a poverty cycle, escape is possible if resources are provided through foreign aid, to be discussed in Chapter 20.

TEST YOUR UNDERSTANDING 19.1

1 a Using a diagram explain the poverty cycle.

b Outline how the poverty cycle is transmitted from generation to generation.

c Explain why external intervention is necessary to break out of this cycle.

19.2 Economic barriers

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain how each of the following works as a barrier to economic growth and development: (AO2)
  - rising economic inequality
  - limited access to infrastructure and appropriate technology
  - low levels of human capital: limited access to education and health care
  - dependence on primary production
  - limited access to international markets
  - informal economy
  - capital flight
  - indebtedness
  - geography and landlocked countries
  - tropical climates and endemic diseases
- evaluate the significance of each of the above factors as a barrier to economic growth and development (AO3)

Economic inequality

The importance of income inequality

Economists have been debating the relationship between economic growth and income distribution over many decades. They used to believe that in the early years of growth, high income inequality is necessary. One reason given was that government spending to alleviate poverty and improve income distribution (such as through spending on merit goods) would take scarce resources away from investments in industrial production, thus reducing growth. Another argument was based on the belief that the rich save more than the poor, resulting in more investment, more physical capital and hence more growth. Therefore, with a more equal income distribution, savings would drop, investments would be lower, and growth would be reduced.

Today, economists argue that the opposite is the case: highly unequal distributions of income are actually *a barrier to growth and development*. *Not only is high income inequality not necessary, but greater equality in income distribution may lead to more rapid growth and more development*. The reasons for this were discussed at length in Chapter 12. It is suggested that you review this section as it is highly relevant to our current topic.

For example, the Asian economies that have grown very rapidly, known as the 'Asian Tigers' (Malaysia, Singapore, South Korea, Taiwan, and others) have more equal income distributions than countries in Latin America and Africa, which have been growing far less rapidly. The Asian Tigers placed a very strong emphasis on the development of human capital, with positive effects on the equality of income distribution and very successful growth and development outcomes.

Rising economic inequality

The performance of developing countries with respect to income distribution varies widely, as in some countries income distribution has improved, whereas in other it has worsened. According to the World Bank, income inequality increased in 2008–2013 in 34 of 83 countries monitored. This was due to more rapid income growth among higher income groups while in 23 countries the poorest 40% experienced a decline in income.[2] In line with global trends inequalities in wealth are worsening in most developing countries. See Chapter 12 for a review of this topic.

Limited access to infrastructure

The importance of infrastructure

Infrastructure is a type of physical capital, and therefore results from investments. It is important for the effective functioning of any economy and includes numerous goods and services such as power, telecommunications, internet access, piped water supplies, sanitation and sewerage, roads, dam and canal works for irrigation and drainage, urban transport, ports and airports.

Infrastructure plays a major role in most economies, representing about 20% of total investments in developing countries. The availability of infrastructure, and broad access to the services it offers make major contributions to economic growth, economic development and poverty alleviation.

Infrastructure provision practically everywhere in the world is mostly a government responsibility. Government's role is the result of a long historical process justified by infrastructure's great political, economic and strategic importance, concerns about market power if it were in private hands, and interest in providing essential services to the overall population of a country. Note that many types of infrastructure qualify as public goods or as merit goods, i.e. goods with positive externalities in consumption.

Barriers to infrastructure development

Most developing countries perform poorly in infrastructure development and provision. For example, in 2015, 1 in 4 health care facilities did not have access to basic water services. 3 in 10 people lacked access to safely managed drinking water while 6 in 10 lacked access to safely managed sanitation facilities. 2.4 billion people did not have access to basic sanitation services like toilets and latrines. 13% of the world's population had no access to electricity.[3]

The reasons for poor performance can be summarised under the following headings:

- **Problems of financing.** Governments charge users for their consumption of infrastructure services (for example, charges for connection to piped water and sewerage, water use, telephone charges, etc.). However,

[2] http://hdr.undp.org/sites/default/files/2016_human_development_report.pdf

[3] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

REAL WORLD FOCUS 19.2

Global warming has increased global economic inequality

A study published in the Proceedings of the National Academy of Science of the United States has found that 'global warming has very likely exacerbated global economic inequality'. It is argued that there is a greater than 90% likelihood that *per capita* GDP in most poor countries is lower than it would have been in the absence of global warming.

> *'Thus, our results show that, in addition to not sharing equally in the direct benefits of fossil fuel use, many poor countries have been significantly harmed by the warming arising from wealthy countries' energy consumption . . .*
>
> *There is growing evidence that poorer countries or individuals are more negatively affected by a changing climate, either because they lack the resources for climate protection or because they tend to reside in warmer regions where additional warming would be detrimental to both productivity and health . . .'*

Figure 19.3: Balaka District, Malawi. Girls carrying buckets of water on their way back from a borehole. Global warming is expected to contribute to water scarcity that may displace up to 700 million people worldwide by 2030 (https://www.unwater.org/water-facts/scarcity)

At the same time, it is possible that some developed countries are benefitting from global warming, and these effects are expected to become more pronounced in the future.

> *'. . . empirical evidence combined with projections of future climate change suggests that, although some wealthy countries in cooler regions could benefit from additional warming, most poor countries are likely to suffer . . .'*
>
> *The relationship between temperature and economic growth means that long-term warming will generally increase growth in cool countries and decrease growth in warm countries. For example, for cooler countries such as Norway, warming moves the country-mean temperature closer to the empirical optimum, resulting in cumulative economic benefits. In contrast, for warm countries such as India, warming moves the country-mean temperature further from the optimum, resulting in cumulative losses.*

The study also notes that:

> *'. . . given that wealthy countries have been responsible for the vast majority of historical greenhouse gas emissions, any clear evidence of inequality in the impacts of the associated climate change raises critical questions of international justice.'*

Source: *www.pnas.org*

Applying your skills

1. Explain why poor countries in warmer climates tend to face heavier consequences of global warming.
2. To what extent should developed countries take responsibility for measures to correct the problems of global warming?
3. Identify a country that has already begun to be affected by global warming. Research the potential consequences.

in developing countries, as part of government social policy intended to make services affordable, the prices charged for infrastructure services have been kept below cost, resulting in insufficient revenue for the state enterprises providing infrastructure.

- **Inadequate maintenance and poor quality.** Lack of revenues means that infrastructure facilities are often poorly maintained, resulting in low quality and unreliable services.
- **Limited access by the poor.** Lack of revenue also means constraints in quantity of infrastructure facilities that can be constructed, so that in many countries it has not been possible to provide services for the entire population. It is generally the poor who tend to suffer disproportionately from lack of access, both in rural areas and in urban slums.
- **Misallocation of resources.** The infrastructure provided is sometimes inappropriate given the needs of the population and the country's level of economic development; investments may be made in infrastructure facilities that remain underused because there is not enough demand for their services (such as power, telecommunications, ports), while other services (such as more roads, or better maintenance and improvements in service quality) are neglected.
- **Neglect of the environment.** Infrastructure can have numerous negative environmental effects, including the failure to adequately control unnecessary emissions, wasteful consumption of water in poorly designed or poorly maintained irrigation facilities, building roads and dams in ecologically vulnerable areas, and others.

REAL WORLD FOCUS 19.3

UN resolution recognises water as a right

On 28 July 2010, the United Nations General Assembly passed a resolution recognising access to clean water and sanitation as a fundamental human right. The resolution states that 'the right to safe and clean drinking water and sanitation [is] a human right that is essential for the full enjoyment of life and all human rights'. It urges governments and international organisations to support efforts to provide safe, clean, accessible and affordable drinking water and sanitation to all.

Figure 19.4: Bishkek, Kyrgyzstan. Gathering drinking water vulnerable to contamination from trash and bacteria causing hepatitis and other water-borne illnesses

Lack of access to clean water and sanitation is one of the most important barriers to economic development and the achievement of decent living standards. According to the Sustainable Development Goals, every day nearly 1000 children die due to water and sanitation-related diarrheal diseases that are preventable. More than 40% of the world's population is affected by water scarcity and this is expected to rise. A mere 27% of the population in developing countries has access to basic handwashing facilities (2015 data).[4]

Water is being increasingly converted into a commodity, which is priced and sold. This weakens the idea of access to water as a human right. The UN resolution was therefore welcomed as a historic step in this process by supporters of water rights.

The Sustainable Development Goals (SDGs) are committed to achieving universal and equitable access to safe and affordable drinking water by 2030.[5]

Source: *www.un.org*

Applying your skills

1 Explain different possible ways that access to clean water can contribute to economic development.

2 Do you agree that access to water should be considered as a fundamental and universal human right? Why?

[4] https://sustainabledevelopment.un.org/sdg6
[5] https://sustainabledevelopment.un.org/sdg6

Limited access to appropriate technology

New technology contributes to improving the *quality* of physical capital. While new technology contributes to economic growth (see Chapter 11), in developing countries it is especially important to consider the *appropriateness* of new technologies to local conditions. There are many technologies developed and used in economically advanced countries that are not well-suited to the conditions of less developed ones. To be effective, technologies must be well-suited to particular economic, geographical, ecological and climate conditions. **Appropriate technology** is a technology that satisfies these conditions.

Different factor supplies (labour and physical capital)

Developing countries are characterised by relatively large quantities of labour (much of which is unemployed or underemployed), while physical capital is relatively scarce and costly to acquire. Moreover, the capacity of developing countries to produce, operate and maintain technologically advanced equipment is limited. In developed countries, labour is relatively scarce in relation to more abundant physical capital, and there is a large capacity for producing and maintaining technologically advanced machines and equipment. Therefore, developing and developed countries have different needs with respect to the kinds of capital goods/technologies that are best suited to their conditions. Developing countries need capital goods and technologies that make use of their abundant labour supplies and that are relatively simple to produce, maintain and operate.

Consider a farm in a less developed country where there is widespread rural unemployment and poverty.[6] Traditionally, farming has been carried out by use of a primitive technology involving the spade, hoe, hand sickle and stick plough. The farm now has the opportunity to introduce a new technology. There are two choices: heavy agricultural machinery, including the tractor, or the use of ploughs pulled by animals. Both these new technologies would increase the productivity of labour. Which one should be adopted?

The tractor technology is more expensive, and uses far less labour to run compared to the previous plough technology. This means many of the workers on the farm will lose their jobs and their income, with rural poverty increasing. Also, it cannot be produced domestically and will have to be imported; this involves the use of scarce foreign exchange with negative effects on the balance of payments. Further, skilled labour will be needed to repair it and parts may have to be imported.

The plough technology is less expensive. It requires a greater quantity of labour for its use, so employment on the farm will increase, leading to less rural unemployment and underemployment, creating more income and reducing rural poverty. It can be produced locally, resulting in increased employment outside the farm as local workers begin to manufacture the equipment. Locally available skills will be used, and more workers can be trained on the job to produce the new equipment. There will be no need for imports and scarce foreign exchange will be saved, with no negative effect on the balance of payments.

Clearly, the farm should adopt the plough technology. We can therefore make a distinction between the following two kinds of technology:

- *Labour-using (labour-intensive) technologies* use more labour in relation to capital. They result in increases in local employment and the use of local skills and materials, increases in incomes and poverty alleviation, and save on the use of scarce foreign exchange.
- *Capital-using (capital-intensive) technologies* use more capital in relation to labour. In developing countries with large supplies of labour they displace workers and increase unemployment, reduce incomes and throw people into poverty, and require skill levels that may be costly and difficult to acquire, as well as the use of foreign exchange for imports.

Most technological advances tend to be of the capital-using type, because most research and technological developments occur in developed countries that focus on their own priorities and needs. This poses serious problems for developing countries that mostly require labour-using technologies and have limited resources for developing technologies well suited to their own economic and physical environments.

[6] See B. F. Johnson and P. Kilby (1975) *Agricultural and Structural Transformation: Economic Strategies in Late-Developing Countries*, Oxford University Press, where these ideas are explored more fully.

Many developing countries have at times tried to copy or imitate the production techniques of the developed world, resulting in the use of *inappropriate technologies*, such as tractor technologies in agriculture, and capital-intensive technologies in industry. Among the failures of growth and development policies, a major one was the push for industrialisation using modern, capital-using technologies in both industry and agriculture. This contributed greatly to unemployment and underemployment and growth of the informal economy in many countries around the world (see Chapter 20).

Different climate and ecological conditions

Many of the technologies developed in rich countries are also inappropriate to the climates, geography and ecological conditions of many developing countries. This is particularly relevant to the needs of agriculture, characterised by different climatic conditions, soils, disease agents and other agronomic factors.

Difficulties in the development of appropriate technologies

Although some developing countries have significantly increased investments in research and development (R&D) (for example, China, India, Brazil), most technological innovation occurs in developed countries. Many developing countries have very few resources to devote to R&D and new technology development. Many developed countries spend more than 2% of GDP on R&D, for example Sweden 3.3%, Japan 3.2%, Denmark and Germany 2.9%, Finland 2.8%, United States 2.7%). By contrast many lower income countries typically spend less than 0.5%.[7] Taking into account their far lower *per capita* GDPs, poor countries spend next to nothing in absolute terms compared to the developed world.

A further problem is that the private sector in developing countries faces few incentives to engage in R&D. In developed countries the private sector is responsible for over 50–60% of R&D expenditures; however, in low-income countries the private sector contribution is very low (from 2% to roughly 20%). Many developed country firms produce and innovate for large markets, and the expectation of large profits creates powerful incentives to innovate in order to compete, capture market shares and take advantage of new opportunities. Firms in developing countries, especially in the lower income ones, have neither the resources nor the markets to support R&D activities.

Low levels of human capital

Human capital refers to skills, abilities and knowledge acquired by people as well as good levels of health, all of which make them more productive. Human capital is important in any society but in developing countries it acquires a special significance because there are large portions of populations in many countries that have relatively low levels of educational attainment, and also low levels of health. In Chapter 18, we saw that developing countries lag behind in both education and health. There is therefore a huge scope for increasing human capital, which can make a significant difference to productivity, employment opportunities, output growth and development prospects.

Education and health, as the two component parts of human capital, are also among the key goals of economic and human development. They are included in the Sustainable Development Goals and are also two of the three dimensions of the Human Development Index (see Chapter 18).

Barriers to education

SDG 4 states 'Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all'. There are several barriers to inclusive and equitable quality education in developing countries.

- **Insufficient funding for education.** Many developing countries, especially the poorer ones, do not have sufficient government revenues to fund education. At the onset of the SDGs there was a $39 billion gap to provide quality education to all children by 2030.[8] Many developing countries should increase the proportion of government budget funds spent on education. As Table 18.5 showed, many countries can do far better with education given their levels of GNI *per capita*. In addition, more aid funds should be directed to education.
- **Insufficient teachers or untrained teachers.** There are not enough teachers at the primary and secondary levels, and those that are available do not have the necessary training. The result is that many children do not receive the basic training they need to learn

[7] The World Bank, World Development Indicators https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS?view=chart

[8] https://www.globalcitizen.org/en/content/10-barriers-to-education-around-the-world-2

THEORY OF KNOWLEDGE 19.1

The value of education

The energy of the mind is the essence of life
Aristotle, ancient Greek philosopher

A mind is a terrible thing to waste
Campaign slogan of the United Negro College Fund in 1972

An investment in knowledge pays the best interest
Benjamin Franklin, one of the Founding Fathers of the United States

He who opens a school door, closes a prison
Victor Hugo, French philosopher

Education is the most powerful weapon which you can use to change the world
Nelson Mandela, South African anti-apartheid activist

Education breeds confidence. Confidence breeds hope. Hope breeds peace
Confucius, Chinese philosopher

Education is a powerful weapon against countless ills. On a personal level it is the means through which an individual can realise her or his full potential as a human being. On a societal level it is the means through which an individual can contribute to the greater social good.

Thinking points

1 Consider each of the above statements. Examine what each statement means regarding the contribution education makes

   a to human development (Chapter 18, Section 18.2), and

   b to economic growth and economic development.

2 Some people in both rich and poor countries may not fully appreciate the value of education. How do you think this issue should be dealt with in a society? Do some people (such as the government) have the right to impose education on other people who may not want themselves or their children to be better educated?

the basic skills of reading, writing and arithmetic. For example, in rural India nearly two-thirds of third grade children cannot solve a 2-digit subtraction problem, and half of fifth graders still cannot do it.[9]

- **Insufficient classrooms and basic facilities.** Many countries in sub-Saharan Africa do not have enough classrooms, and many of those that are available do not have running water toilets. For example in Chad only one in seven schools has potable water and one in four has toilets, while only one-third of available toilets are for girls only.[10]
- **Lack of teaching materials.** Textbooks are often old and worn out, and those that are available are shared by several children. In Cameroon for example, there are 11 primary school students for every reading book, and 13 second grade students for every arithmetic book. Similarly all other teaching materials are in very short supply.[11]
- **Children with disabilities are excluded.** In the poorest countries it is estimated that 95% of children with disabilities do not attend school while also facing discrimination.

[9] https://www.worldbank.org/en/news/immersive-story/2019/01/22/pass-or-fail-how-can-the-world-do-its-homework
[10] https://www.globalcitizen.org/en/content/10-barriers-to-education-around-the-world-2
[11] https://www.globalcitizen.org/en/content/10-barriers-to-education-around-the-world-2/

- **Gender discrimination.** It is estimated that over 130 million girls do not go to school on account of being a girl (see also Table 19.4 below).
- **Conflict or risk of conflict.** It is estimated that 250 million children live in countries where there are conflicts, while 61 million do not go to school because of conflicts.
- **Distance of school from home.** It is not uncommon for a child to have to walk three hours a day to get to school.
- **Hunger and malnutrition.** Many millions of children are malnourished. Being malnourished seriously affects a child's ability to learn.
- **Inability to pay for education.** Whereas in principle primary education is free virtually everywhere, expenses of having to buy books, pens, exam fees or other 'informal fees' means that many families locked in poverty cannot afford to send their children to school.

Barriers to achieving good health

SDG 3 states 'Ensure healthy lives and promote well-being for all at all ages'. Barriers include:

- **Insufficient funding for health care.** It is estimated that there is a financing gap of $370 billion in order the reach the SDG targets of goal 3 noted above. The financing gap is especially serious in sub-Saharan Africa which makes up 16% of the world's population, but carries 23% of the world's disease burden, while accounting for 1% of global health care spending.[12]
- **Insufficient access to health care services.** A study by the World Bank and World Health Organization has found that at least half of the world's population do not have access to all the essential health care services they need.
- **Private payments for health care.** The World Bank and World Health Organization estimate that 800 million people spend at least 10% of their budget on health care services, while 100 million are pushed into extreme poverty (living on less than $1.90) a day because of this.[13] For sub-Saharan Africa it is estimated that 37% of all spending on health care is made out-of-pocket. This has disastrous effects as 11% of the African population is pushed into poverty as a result of having to make these payments while another 37% does not receive medical treatment due to the high cost.[14]
- **Geographical access.** Health care facilities are often situated very far from where people live. Long distances, lack of roads and transportation mean that many people do not have access to health care facilities. Research has found that standards of health care improve when distance is reduced. For example, in Ghana it was found that when the distance to a health care facility was halved the utilisation of the services doubled.[15]
- **Insufficient numbers of trained medical practitioners.** This is an extremely serious problem because even where medical practitioners are available they are often ineffective. The World Health Organization estimated for seven African countries that medical practitioners make correct diagnoses one-third to three-fourths of the time.[16]
- **Insufficient medical facilities and medical supplies**. There are major problems of financing in order to set up, staff and equip medical facilities to ensure access to rural populations.
- **Acceptability of modern medical practices.** In some very poor societies modern medical practices are not accepted as people prefer traditional medicine.
- **Insufficient access to clean water and sanitation.** Adequate access to quality health care services is not enough to ensure good health. It is also necessary to ensure access to clean water supplies and sanitation. For example, there are more than 340 000 children per year who die due to diarrhoeal diseases from unsafe drinking water and poor sanitation; this is entirely preventable.[17]

[12] https://www.brookings.edu/blog/future-development/2019/03/01/closing-africas-health-financing-gap/

[13] https://www.who.int/news-room/detail/13-12-2017-world-bank-and-who-half-the-world-lacks-access-to-essential-health-services-100-million-still-pushed-into-extreme-poverty-because-of-health-expenses

[14] https://www.brookings.edu/blog/future-development/2019/03/01/closing-africas-health-financing-gap

[15] http://www.hhpronline.org/articles/2016/11/10/healthcare-in-the-remote-developing-world-why-healthcare-is-inaccessible-and-strategies-towards-improving-current-healthcare-models

[16] https://www.who.int/news-room/detail/05-07-2018-low-quality-healthcare-is-increasing-the-burden-of-illness-and-health-costs-globally

[17] WaterFacts_water_sanitation_hygiene_Sep2018.pdf

TEST YOUR UNDERSTANDING 19.2

1 Use *PPC* diagrams and *AD-AS* diagrams to explain the likely effects of

   a greater economic equality,

   b the development of new infrastructure,

   c the use of appropriate technology, and

   d improvements in human capital.

Dependence of production and exports on the primary sector

The **primary sector** of an economy is the sector that produces *primary commodities*, which are goods arising from the factor of production *land* (see Chapter 1). They include agricultural, fishing and forestry products, as well as products of extractive industries (oil, coal, minerals, etc.). Developing countries, especially the lower income ones, tend to specialise in the production of only a few goods, which usually are primary commodities. The exports of any country are determined to a large extent by the goods it produces. Therefore, if it specialises in primary products, its exports are likely to be dominated by primary products.

REAL WORLD FOCUS 19.4

Education and health as fundamental human rights

There is a general belief around the world that education and health are fundamental human rights. According to this view, everyone should be entitled to receive an education and to have access to services allowing them to achieve their potential for good health. This principle is reflected in the United Nations Universal Declaration of Human Rights (1948), affirming the individual's right to 'a standard of living adequate for the health and well-being of himself and his family, including food, clothing, housing and medical care' as well as a right to education that is compulsory and 'free, at least in the elementary and fundamental stages'. The rights to education and health have been further asserted by the International Covenant on Economic and Social Rights (1966), and are guaranteed by the national constitutions of many countries.

Figure 19.5: Dental health checkup

As fundamental human rights, education and health care are crucially important ingredients in the process of expanding human freedoms, the defining characteristic of human development (see Chapter 18, Section 18.2). An individual who is literate and knowledgeable, adequately fed, free of preventable diseases, and who has access to clean water, sanitation and health care services, has significantly improved opportunities to find employment, have a decent income, escape extreme poverty, and lead a long and healthy, creative life and to enjoy a decent standard of living, freedom, dignity, self-esteem and the respect of others.[18] Evidence accumulated throughout the world overwhelmingly shows that education leads to improved employment opportunities and higher levels of income. Improved health reduces illness and physical disability, increases life expectancy, improves quality of life and reduces poverty.

Applying your skills

Based also on the discussion in Chapter 20, Section 20.4, examine economic and non-economic reasons that justify government intervention in the provision of education and health care. Explain which of these you consider to be the most important.

[18] United Nations Development Programme, *Human Development Report 1997* (available at http://hdr.undp.org/en/reports)

Country	Commodities as a % share of total exports	Leading commodity export as a % share of total commodity exports	
		Commodity	Share in commodity exports %
Bolivia	95	natural gas	48
Botswana	94	pearls, precious and semi-precious stones	85
Central African Republic	90	forestry products	41
Chad	98	oil	92
Chile	86	copper	32
Comoros	77	spices	91
Cote d'Ivoire	86	cocoa	45
Gabon	95	oil	79
Malawi	84	tobacco	60
Mali	92	gold	71
Mongolia	98	copper	49
Nigeria	97	oil	78
Seychelles	90	fishery products	84
Somalia	95	live animals	70
Venezuela	92	oil	81
Zambia	86	copper	79
Zimbabwe	83	tobacco	40

Source: *UNCTAD State of Commodity Dependence 2016*

Table 19.1: Export dependence on primary commodities

Many of the countries in sub-Saharan Africa and Latin America receive at least half their export earnings from primary commodities. Table 19.1 presents a list of selected developing countries that are *overspecialised*: they are highly dependent on commodity exports, as well as on a single product that represents a high percentage of the total value of commodity exports. Countries such as these face a number of risks and obstacles that work against growth and development objectives, discussed below.

Price volatility of primary products

This topic was explained in Chapter 3, where we saw that the prices of primary products are more *volatile* (fluctuate more) than the prices of manufactured products because they have low price elasticities of demand and low price elasticities of supply (both are usually less than one). We also examined why primary products have low *PED*s and *PES*s.

Volatility of primary product prices has serious negative consequences for producers and for the economy as a whole. As product prices fluctuate, so do farmers' incomes, along with agricultural investment, employment and wages of agricultural workers. Producers are unable to plan, as they are unable to determine the future profitability of investments.

When the product is exported, fluctuating prices also translate into fluctuating and unstable export earnings, affecting the country's balance of payments and ability to import. In addition the price fluctuations affect the government's revenues with negative consequences on its efforts to plan for growth and development and undertake necessary investments in merit goods (health care, education infrastructure).

For these reasons, volatility of primary product prices has been a major barrier to growth and development in many developing countries that are highly dependent on a few primary products for their export earnings.

Limited access to international markets

The inability to access international markets refers to difficulties encountered by developing countries in their exports to developed countries.

Tariff barriers

- ***Products of interest to developing countries face relatively high tariff barriers***

There are important differences regarding tariff levels between products and product categories. In particular, tariffs remain high for products of interest to developing countries, which include agricultural products, textiles and apparel.[19] Agricultural products in particular face much higher tariffs than those in manufacturing and natural resources.

- ***Use of tariff barriers to discourage the development of manufacturing and diversification***

When countries specialise in primary products or other raw materials used for manufacturing, the domestic availability of the raw materials works to stimulate manufacturing in products based on these raw materials. It is much easier for an economy to produce chocolate when it produces the cocoa beans that are the basic input for chocolate. This is called *vertical diversification*, and involves moving into manufacturing that makes use of domestically produced inputs. A number of developing countries provided a major boost to their growth and development through this type of diversification (for example, Malaysia, Thailand, Indonesia, China, Chile and Mauritius).

However, both developing and developed countries impose low tariffs on raw materials (unprocessed primary products), and much higher tariffs on processed products. This is termed *tariff escalation*. Tariff escalation makes it difficult for developing countries to expand into manufacturing, since the greater the degree of manufacturing, the more difficult it is for the product to be exported to developed countries.

Clearly, tariff escalation works to discourage developing countries from diversifying their production into manufacturing. Tariff escalation occurs most often in products that are important to developing countries, such as apparel, animal products, tanning and manufacturing.[20]

Agricultural trade and rich country subsidies

Agricultural support by rich countries is one of the most problematic issues in international trade. Agriculture is one of the most protected sectors of developed countries, justified on the grounds that farmers earn low incomes and must therefore be assisted. Yet the evidence indicates that most of the benefits of protection are actually enjoyed by very large farmers.[21] This protection has major negative consequences for many developing country primary product exports and for poverty alleviation.

Most developed countries have some form of support system for their farmers, of which the two best known are the European Union's Common Agricultural Policy (CAP) and the United States' farm policy. Both farm protection systems involve a mix of price floors and subsidies (see Chapter 4).

Negative consequences of developed country farm support in developing countries

Over and above the negative consequences of protection in the domestic economy (high consumer prices, protection of inefficient producers, costs to taxpayers of protection, etc. see Chapter 4), there are a number of negative consequences for the global economy and developing countries:

- **Global misallocation of resources.** Higher prices received by developed country farmers due to price supports, as well as production subsidies, result in an overallocation of resources to the production of protected goods in the developed countries, and therefore excess production and surpluses. When these goods are exported they artificially lower the international price of the goods, making it more difficult for farmers in developing countries to compete. Very low prices force some farmers in developing countries to abandon or reduce cultivation of the product. Therefore too much of

[19] https://unctad.org/en/PublicationsLibrary/ditctab2019d1_en.pdf

[20] https://unctad.org/en/PublicationsLibrary/ditctab2019d1_en.pdf

[21] It is not surprising that the largest farmers get most of the benefits of protection, since subsidies are paid out on the basis of the size of the farm, as well as on the basis of historical production, or quantity produced over a period of years.

the protected good is produced in the developed countries (overallocation of resources) and too little in developing countries (underallocation of resources).

- **Global inefficiency.** Developing countries can often produce certain agricultural products at a far lower cost than developed countries, yet because of protection, the more inefficient developed country producers continue to produce, capturing global market shares from the more efficient developing country producers. For example, the United States is the world's largest exporter of cotton, which receives price supports and subsidies worth billions of dollars annually, yet it also has among the highest costs of cotton production in the world.
- **Lower export earnings for developing countries.** Developing countries that export products receiving protection in developed countries suffer due to lower exports, as well as lower prices (due to the rich-country subsidies), and therefore have lower export earnings; this can contribute to balance of payments difficulties and increased debt burdens.
- **Increased poverty among affected farmers.** Low exports and low prices received by developing country farmers and the inability to compete with the farmers of developed countries means lower incomes, lower investment possibilities for farmers, lower employment opportunities for farm workers and increased poverty. Some displaced farmers are forced to migrate to the cities, where they find work in the informal economy while contributing to the growth of urban slums.

Other non-tariff barriers: the 'new trade protection'

There is some concern that certain non-tariff barriers have begun to increase in recent years, the most important of which involve technical regulations, standards and requirements, testing and certification, labelling and packaging requirements; customs and administrative procedures; and sanitary measures including food safety and quality standards. These are referred to as *administrative barriers* in Chapter 14. (See Real world focus 15.1 in Chapter 15.) Because these kinds of barriers have been rising rapidly in recent years, they are referred to as the 'new trade protection'. Non-tariff barriers cover the entire range of products exported by developing countries. Whereas the need for minimum safety and quality standards is not in question, there are concerns that some of these controls are being used as a hidden way to limit quantities of imports.

TEST YOUR UNDERSTANDING 19.3

1. Explain why too much specialisation on primary commodity production and exports may have negative effects on developing countries' efforts to grow and develop.
2. Examine the possible negative effects of short-term volatility of commodity prices.
3. Identify some factors that make it difficult for developing countries to access foreign markets for their exports.

The informal economy

A *formal economy* refers to the part of an economy that is registered and legally regulated; an **informal economy** by definition lies outside the formal economy, and refers to economic activities that are unregistered and legally unregulated.

Informal economies exist everywhere in the world, but are much more important in developing countries. In developed countries, the informal economy includes unregistered work resulting in tax evasion, as well as corruption or crime which are illegal. In developing countries, the informal economy offers work that can make all the difference between physical survival and starvation for individuals and their family. As the International Labour Organization (ILO)[22] stresses, work in the informal economy in developing countries is not undertaken to avoid payment of taxes or bypass labour or other legislation as in developed countries. 'ILO constituents from countries all over the world agreed that most people enter the informal economy not by choice, but as a consequence of a lack of opportunities in the formal economy and in the absence of other means of livelihood.'[23]

Examples of activities in the informal sector in urban areas include everything from barbers, cobblers, carpenters, tricycle and pedicab drivers, garbage collectors and small shop owners, to working in restaurants, hotels and sweatshops (manufacturing enterprises where

[22] The International Labour Organization is an agency of the United Nations concerned with global labour issues.
[23] p.49 https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_626831.pdf

Region	Share of informal employment in total employment, % 2016
Africa	85.8
Arab States	68.6
Asia and Pacific	68.2
Americas	40.0
Europe and Central Asia	25.1
Total	61.2

Source: ILO, Women and Men in the Informal Economy: A statistical picture, 2018

Table 19.2: Size of the informal economy globally

workers are paid very low wages and work long hours under unhealthy conditions), in construction, in domestic household work or in offices as temporary help. In rural areas where the informal economy is even more important, many activities relate to agriculture.

A major study by the ILO estimates that two billion people aged 15 and above, corresponding to 61.2% of global employment, work in the informal economy.[24] Of these two billion people, 93% are in middle-income and low-income developing countries. Table 19.2 shows the breakdown by geographical region.

The level of informal employment is closely linked with a country's level of human development as measured by the Human Development Index (HDI). Countries with a high level of informal employment have a lower HDI value. This is because the level of informal employment is greater in lower income countries, and it is also affected by levels of education. Informal employment decreases as the level of education increases. People who have completed secondary or tertiary (university) education are less likely to be in the informal economy compared with workers with no education or primary level education.

In addition, people in rural areas are nearly twice as likely to have informal employment compared to people in urban areas.

The informal economy poses many problems. Clearly if workers are unregistered the government loses tax revenues. But in addition and very importantly, there are serious problems for the workers themselves as there is no worker protection. Workers are vulnerable to exploitation; environmental dangers and health hazards in slums with no basic services like water sanitation and sewerage; no access to credit for workers; limited possibilities for education and training; no social protection including pensions, and many more.

Goal 8 of the Sustainable Development Goals, which is to 'Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all' includes a target and indicator intended to tackle the problem of the informal economy. According to one of the authors of the ILO study noted above

> *'There is an urgent need to tackle informality. For hundreds of millions of workers, informality means a lack of social protection, rights at work and decent working conditions, and for enterprises it means low productivity and lack of access to finance.'*

The Director of ILO's Department of Statistics adds

> *'The high incidence of informality in all its forms has multiple adverse consequences for workers, enterprises and societies and is, in particular, a major challenge for the realisation of decent work for all and sustainable and inclusive development. Having managed to measure this important dimension, now included in the SDG indicators framework, this can be seen as an excellent step towards acting on it, particularly thanks to more available comparable data from countries.'*[25]

Capital flight

Capital flight refers to the large-scale transfer of privately-owned financial capital (funds) to another country. It should be distinguished from normal financial capital outflows that occur because of the desire of residents to diversify their holdings of foreign assets. Instead, it results from high uncertainty and risk associated with holding domestic assets, and occurs when the residents or businesses of a country are fearful that their wealth or income or livelihood (or even their physical safety in extreme cases) are being threatened in the country of residence. They may be fearful about the prospect of loss of property through confiscation, or sudden increases in taxation, or political instability that may lead to social and economic turmoil with negative effects on economic performance, or about anything that may lead to the loss of value of the domestic currency, such as high rates of inflation, serious balance

[24] https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_626831.pdf
[25] https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_627189/lang--en/index.htm

of payments problems, serious foreign debt problems, and the possibility of devaluation or depreciation of the domestic currency. In all these cases financial capital leaves the country in search of safe investments in assets in a safer financial, economic and political environment, and the domestic currency is exchanged for stable currencies that are not expected to lose their value.

Capital flight can be a major problem, because it involves a loss of financial capital that could have been invested domestically. Further, as it consists of the sale of the domestic currency (in order to purchase foreign exchange that leaves the country), it exerts a downward pressure on the value of the currency, often forcing governments to devalue or allow the currency to depreciate. When capital flight is prompted by fears of currency devaluation, it has the effect of becoming a self-fulfilling prophesy: fear of devaluation leads to sales of the domestic currency, which cause the currency to be devalued or depreciated. Sometimes, due to the impacts of capital flight, a planned devaluation is carried out earlier or is larger than what the government had originally envisioned.

Capital flight has played a major role in worsening the external debt problems of many developing countries, particularly during the 1980s (see the discussion on debt below). It involves the use of scarce foreign exchange, and it is the scarcity of foreign exchange that had led many countries to borrow increasingly. Capital flight therefore increases the need for external borrowing. It has been estimated that in the period 1973–1987, the amount of capital flight in several Latin American countries accounted for more than half the increase in foreign debt over the same period. In Argentina 61% and in Mexico 64% of foreign debt accumulation was due to capital flight. In Venezuela, capital outflows were even greater than foreign debt accumulation; in other words the funds flowing out due to capital flight were greater than the funds flowing inward due to new loans![26] Venezuela is experiencing similar problems in more recent years.

Capital flight has also played a major role in financial crises. In Mexico, political instability (political assassinations and guerilla rebellions) and lack of confidence in the economy led to massive sales of pesos (the Mexican currency), a huge drop in foreign exchange reserves, massive capital flight, peso devaluation, and a fall in real output of 7% in 1994–1995. In Russia, in the late 1990s, loss of confidence in the ruble (the Russian currency) and in the government's ability to repay its debt led to fears of a ruble devaluation and to massive flight from rubles into foreign currencies. Interest rates paid by the government on its ruble loans climbed to 150%, and in spite of IMF and World Bank loans intended to support the ruble, it was devalued in 1998, and lost nearly half its value in the course of a few months. In several East Asian countries, in the late 1990s, capital flight played a major role in precipitating a crisis that threatened global financial stability.

Indebtedness

The beginnings of the debt problem date back to the oil shock of 1973–1974, when the Organization of the Petroleum Exporting Countries (OPEC) suddenly increased the price of oil. Almost overnight, oil-importing developing countries were faced with larger import expenditures due to higher oil prices and began to borrow to cover their deficits. By the early 1980s many developing countries had amassed very high levels of debt and were on the verge of bankruptcy.

Historical background (Supplementary material)

If you would like to read more about how the debt problem emerged and developed, you may do so in the digital coursebook.

More recent developments

The difficulties caused by high levels of debt led to pressure on creditors to cancel debts of highly indebted countries. These initiatives will be further discussed in Chapter 20.

In 1996, the World Bank and IMF began the Heavily Indebted Poor Countries (HIPC) Initiative, intended to provide **debt relief** to some highly indebted poor countries by cancelling a portion of their debts. Since then, 30 African countries received over $100 billion in debt relief, which allowed them to reduce their debt service costs (explained in Chapter 11) and increase spending on social services with significant positive effects: reduction of poverty and hunger, and increased spending in primary education and health, with highly favourable effects on health indicators (reduction of infant and child mortality rates, incidence of tuberculosis and HIV/AIDS).

However, the debt situation of many poor countries has again deteriorated. By 2018, more than one-third of the countries that had received debt relief were back or nearly back at their pre-HIPC debt levels, while only 4 of the 30 countries were at low risk.

[26] Economic Policy Institute, (1987) *Capital Flight and the Latin American Debt Crisis*, Washington, D.C.

There are several factors behind these developments.[27]

- The global financial crisis (GFC) that began in 2008 hit many economies around the world, resulting in economic downturn and recession. Before this time, most of the countries in question had a budget surplus, but with the onset of the GFC the surplus turned into deficit.
- In 2014, there was a major drop in global commodity prices. As all the countries in question are commodity exporters, they suffered a great loss of export revenues resulting in even larger deficits. You may remember that budget deficits are covered by borrowing, therefore the increasing deficits led to the need for more borrowing.
- Interest rates in economically developed countries dropped to very low levels as central banks followed very easy monetary policies in order to address the recessions caused by the GFC. This made it easier for some African countries to borrow from international lenders.
- Low levels of saving and large financing needs for infrastructure in African countries led them to borrow more in order to invest in infrastructure projects. China has become a major lender to sub-Saharan African countries, particularly for infrastructure projects (roads, rail) having spent $125 billion on loans in the period 2000–2017.
- Poor governance (to be discussed below) led to an increase in debt.
- Exchange rate depreciation following the commodity price decreases caused the real value of debt to increase (see Chapter 16).
- In a number of countries increased borrowing led to more government spending rather than investment
- There was an increase in borrowing from private sources (commercial banks) which lend at market rates (that are higher than World Bank and IMF rates for very poor countries), resulting in higher debt-servicing costs.

Rising debt levels are creating concerns that in some countries debt may be increasing to unsustainable levels, with serious negative consequences for their growth and development prospects. The costs of high levels of debt were discussed in Chapter 11 (at HL). They will be briefly listed here.

- Debt servicing costs mean that the government has fewer resources for social services, infrastructure, etc.
- Poor credit ratings mean it is increasingly difficult for governments to borrow.
- It may be necessary to increases taxes and cut government spending in order to reduce the level of debt, which is contractionary for the economy.
- There may be increased income inequality if the government issues bonds that are held by higher income people.
- There is lower private investment due to uncertainty.
- There is a possibility of a debt trap where the government keeps borrowing more to pay back its old debts.
- There is lower economic growth due to all of the above.

Geography and landlocked countries

Adam Smith in the 18th century (see Chapter 1) observed that geography plays a role in the development of markets, noting that transportation by sea is less expensive than by land, something that remains true today in spite of huge advances in means of land transport.

Landlocked developing countries (LLDCs) are particularly disadvantaged since in order to access ports for their export and import activities they depend on their neighbouring countries, with which they may sometimes have poor relations, or which may have poor road infrastructure, or which may face conflict and political instability. Even in the best of circumstances they face administrative procedures and high transport costs, which at times may be higher than the value of the output they wish to sell. These factors prevent their integration into the global grading system. In addition, many of these countries are overly dependent on a small range of primary commodities (see above), increasing the range of challenges they face.

Table 19.3 provides a list of landlocked developing countries. Many of these have a low share of world exports, very low GNI *per capita* and low Human Development Index ranks and values.

In recognition of the special problems of LLDCs, the international community has adopted the 'Vienna Programme of Action' for the period 2014–2024[28] with the goal to assist these countries in their efforts to achieve the Sustainable Development Goals (SDGs).

[27] Brookings https://www.brookings.edu/wp-content/uploads/2019/04/africa_sovereign_debt_sustainability.pdf

[28] https://unohrlls.org/custom-content/uploads/2015/03/Vienna-Programme-of-Action.pdf

Africa		Asia	Europe	Latin America
Botswana	Niger	Afghanistan	Armenia	Bolivia
Burkina Faso	Rwanda	Bhutan	Azerbaijan	Paraguay
Burundi	South Sudan	Kazakhstan	Moldova	
Central African Republic	Uganda	Kyrgyzstan	North Macedonia	
Chad	Zambia	Laos		
Eswatini	Zimbabwe	Mongolia		
Ethiopia		Nepal		
Lesotho		Tajikistan		
Malawi		Turkmenistan		
Mali		Uzbekistan		

Table 19.3: Landlocked developing countries

Tropical climates and endemic diseases

Countries differ in their climate, which is an important factor determining the nature of their economic activities. Climate differences are key in determining types and methods of agricultural production, animal husbandry and even labour productivity. For example, heat and humidity may reduce labour productivity, while tropical and subtropical climates are known to reduce soil quality and negatively affect the health of both humans and animals.

While there are obviously climate differences from country to country, on the whole most developed countries have temperate climates, while almost all developing countries have tropical and subtropical climates. This is considered to be a factor that has contributed to the different development patterns of developed and developing countries.

In addition, tropical and sub tropical climates are conducive to certain diseases such as malaria and parasitic diseases (infectious diseases caused by parasites). *Endemic diseases* are those that are very commonly found in an area. For example, in Africa, very common diseases include pneumonia, malaria, diarrhoea and tuberculosis all of which are for the most part curable or preventable. In addition, a major health issue in Africa is HIV/AIDS which is also preventable.

It should be emphasised that most diseases in developing countries are rooted in the consequences of poverty, including poor nutrition, inadequate vaccination coverage, lack of clean water supplies, limited or no access to sanitation, indoor air pollution due to inappropriate cooking fuels, limited access to health care facilities and medications, and poor health education. Most of the diseases in low income countries are either preventable or treatable.

TEST YOUR UNDERSTANDING 19.4

1 Explain why each of the following is a barrier to developing countries' efforts to grow and develop:

- a the informal economy
- b capital flight
- c indebtedness
- d the geography of landlocked countries
- e tropical climates and endemic diseases.

19.3 Political and social barriers

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain how each of the following works as a barrier to economic growth and development (AO2)

CONTINUED

- weak institutional framework
  - legal system
  - ineffective taxation system
  - banking system
  - property rights
- gender inequality
- lack of good governance and corruption
- unequal political power and status
- evaluate the significance of each of the above factors as a barrier to economic growth and development (AO3)

Weak institutional framework

The World Bank defines institutions as 'the rules, organisations and social norms that facilitate co-ordination of human action'.[29] There are many economic, legal and social institutions that influence economic growth. In many economically less developed countries, there is a need to develop institutions relating to property rights (laws and regulations that define rights to ownership, use and transfer of property); a well-functioning legal system that provides effective enforcement of laws, contracts and mechanisms for settling conflicts; an efficient, fair and transparent tax system; banking and credit institutions that provide effective links between savers and investors, and broad access by the population (including the poor) to credit; institutions that protect against corruption; and more. Many developing countries are making great efforts to build institutions since a market system cannot function well without well-developed institutions such as these.

Legal framework and access to justice

A well-functioning legal system is very important to growth and development. This includes equal access to justice for all, including vulnerable groups. According to the United Nations Declaration of the High-level Meeting on the Rule of Law, member states have committed themselves to take 'all necessary steps to provide fair, transparent, effective, non-discriminatory and accountable services that promote access to justice for all'.[30] Access to justice is necessary to eliminate poverty, as poverty is often the result of disempowerment, exclusion and discrimination.

In order to have equal access to justice for all, it is necessary to have the appropriate institutions in place that provide a legal framework for the proper functioning of the economy, including effective tax collection systems and protection of property rights (see below) as well effective ability to resolve disputes.

Ineffective taxation structures

Governments need revenues to be able to make investments in human capital, provide infrastructure, and undertake activities required for growth and development (and other) purposes. Government revenues generally arise from taxation. Yet developing countries often have inadequate taxation institutions and tax structures, resulting in low levels of revenues, often with negative impacts on equity and resource allocation.

Developing country tax systems typically display the following features:

- high dependence on indirect taxation (see Section 12.5, Chapter 12)
- inefficient and highly bureaucratic tax systems
- weak tax collection systems, often marked by a significant level of corruption
- concentration of political and economic power in wealthy groups, enabling them to influence government tax policy so as to minimise their tax burden.

We will now examine the impacts of these characteristics on:

- levels of revenues
- equity
- resource allocation.

[29] The World Bank (2003) *World Development Report, 2003: Sustainable Development in a Dynamic World: Transforming Institutions, Growth and Quality of Life*, Oxford University Press.

[30] https://www.un.org/ruleoflaw/thematic-areas/access-to-justice-and-rule-of-law-institutions/access-to-justice/

Low levels of revenues

International comparative studies show that the lower the GDP *per capita* of a country, the lower the tax revenues as a share of GDP; the higher the level of GDP *per capita*, the higher the tax revenues as a share of GDP. While it is true that poor countries do not have the capacity to generate as much tax revenue as a share of GDP as rich countries (due to low levels of income for substantial portions of the population, which mean that most of income must go to satisfy consumption needs), there are also indications that *tax revenues are lost on account of several factors*:

- corruption in tax collection, involving paying tax collectors bribes to lower the amount of tax paid
- inefficiencies in tax collection, involving complicated bureaucratic procedures and complex tax legislation, that result in waste of resources and also create significant opportunities for tax evasion (avoidance of tax payment)
- tax exemptions and privileges of wealthy people and large firms, made possible by their economic and political power and influence on government policy, or by government efforts to promote particular private sector activities,[31] which are often exploited to the extreme with the result that there are significant losses of tax revenues
- low property tax rates; property taxes could be a significant source of tax revenues in many countries, however property tax rates tend to be much lower in developing countries compared to developed ones. Property taxes quite obviously tax the relatively wealthier groups, who are the owners of property. The problem is particularly acute in countries where the distribution of land ownership is highly unequal (for example in Latin America), where because of the political and economic power of wealthy landowners, governments hesitate to increase tax rates or to enforce payment of taxes
- the political and economic power of elites influence government tax policies generally (not only in connection with property taxes), in order to minimise their tax burden; wealthy, elite groups can afford to purchase privately many merit goods (health services, education, etc.) and usually already have access to the limited levels of infrastructure made available by the government, as these groups are the first to benefit from the provision of publicly provided services (clean water supplies, sewerage systems, communications, etc). Therefore they tend to be far less concerned about payment of taxes for the purpose of enabling the government to undertake investments in services that will benefit the poor. Since higher income groups have a far greater ability to pay taxes than lower income groups, a low tax burden on the rich translates into very significant losses of tax revenues.

Lower levels of tax revenues mean that the government has fewer revenues with which it can undertake necessary investments in essential services (education, health care, water supplies, sanitation, irrigation, road systems, etc.).

Inequities in tax systems

Taxation systems in less developed countries are often inequitable; reasons include the following:

- They are not particularly progressive, and may even be regressive (lower income people pay a higher proportion of income in taxes than higher income people); reasons for the lack of progressivity are that
    - developing country tax systems typically rely far more on indirect rather than direct taxes, including tariffs as compared with more developed countries, because these kinds of taxes are much easier to collect
    - to the extent that tax systems rely on direct taxes, these are not always very progressive.
- The factors noted earlier, including tax evasion, tax concessions and privileges for higher income groups, and low property taxes, mean that very high income groups end up paying far less overall taxes in relation to their income than medium and lower income groups, thereby making the tax system even more regressive.

Inequities in the tax system accentuate income inequalities: there is limited redistribution from higher income to lower income groups.

Negative impacts on resource allocation

All tax systems have impacts on resource allocation by changing relative prices and incentive structures. A well-designed tax system tries to be as neutral as possible with respect to its impacts on resource allocation.

[31] Governments in less developed countries often use tax concessions as a means of stimulating activity by creating incentives for businesses; the beneficiaries of these tax concessions are often multinational corporations, as well as large domestic firms in sectors that the government wants to promote as part of its growth strategy.

In developing countries, taxes often affect the allocation of resources in ways that are far from neutral, and result in worsening, or 'distorting' the allocation of resources. Some examples of poorly designed tax policies with distorting effects include the following:

- many firms or sectors benefit from a variety of tax advantages, such as lower tax rates, which derive from government interests in promoting particular sectors, or because politicians favor allies and supporters; state enterprises often pay little or no tax at all
- medium-sized firms tend to be penalised by having to shoulder a higher tax burden; this occurs because large firms can more easily evade taxes through loopholes and tax advantages
- tax advantages enjoyed by older firms that are well connected with politicians act as a disincentive for new firms trying to enter a market and start up a new business
- imports of capital goods are often exempted from import tariffs; this produces an artificially lower price for capital goods and promotes the use of capital-intensive technologies.

These factors contribute to creating patterns of resource allocation that are not in society's best interests. When resources are misallocated (or poorly allocated), the 'what/how much to produce' and the 'how to produce' questions of economics are not being answered in the best possible way; societies are worse off because the 'wrong' combinations of goods and services end up being produced, and the 'wrong' combinations of resources are used for production (such as use of capital intensive technologies).

Banking

Banking services and access to credit are very important to economic growth and development. They provide the link between savers and investors, making funds saved by savers available to borrowers who wish to make investments. If funds of savers could not be borrowed by investors, people would have to rely on individual savings for investment, and these would hardly be enough to finance the needed levels of new capital formation.

Importance of banking for growth and development

Banking and credit institutions contribute to growth and development in the following ways:

- They provide an incentive for people to save, because they offer a return (in the form of interest) on savings. The greater the savings in an economy, the greater are the funds available to be invested.
- They provide businesses and farmers with credit to open, run and expand their businesses and farms. Increased borrowing permits greater investment, resulting in increased output and therefore growth.
- They provide consumers with credit that can be used for investments in human capital (education for their children as well as essential medical care), increasing the productivity of labour and contributing to growth and development.
- Access to credit is very important for poverty alleviation because poor people are least able to save any part of their income, since they are forced to spend most of it on essentials (such as in the poverty cycle). Making credit available to very low-income earners can therefore allow them to make necessary investments in physical, human and natural capital. It can also contribute to improving the distribution of income as the investments that credit makes possible increase the incomes of the poor and allow them to participate in economic growth.

Yet the commercial banking system in developing countries is not well developed. The total amount of funds in bank deposits on average is about 15% of national income, compared to 30% or more in developed countries. The number of banks and their branches in relation to the population in developing countries is far smaller than in developed ones, so that businesses and consumers may not have easy access to banking services.

Also, there is a high degree of public sector ownership of commercial banks (in contrast to mostly private sector ownership in more developed countries). High public sector ownership has historically been justified on the grounds that this allows governments to guide bank lending and resource allocation to highly productive investments. However, the evidence suggests that public ownership of banks leads to lower levels of efficiency, less borrowing and saving, highly bureaucratic procedures that discourage private investment through borrowing, less competition and lower productivity.

Exclusion of the poor from access to credit

Commercial banks in developing countries often cater to larger, wealthier borrowers, bypassing the medium-sized and especially smaller borrowers. Many commercial banks are overseas branches of large multinational banks that are more interested in providing loans to multinational corporations or large domestic firms. Even domestic banks, often under government direction, prefer borrowers in manufacturing who are considered to be more creditworthy and who are believed to be the driving force of industrialisation. Sometimes, bank lending is influenced by political interests and the power of local elites with large control over the business and financial sectors. Furthermore, commercial banks are interested in large loans that are safe, secured with collateral or some form of wealth. Poor and small-scale producers, farmers and traders need very small loans, and very often lack collateral to secure their loans. Therefore, lack of wealth typically means lack of access to credit.

The result is that the overwhelming majority of people in developing countries are forced to rely on informal (illegal) sources of credit consisting of moneylenders, also known as 'loan sharks', who charge very high interest rates; pawnbrokers who lend amounts far smaller than the value of the article that has been pawned; and friends and family, from whom the supply of loanable funds is irregular or insufficient. It is estimated that almost three billion poor people do not have access to basic financial services that are essential to managing their lives.[32]

Lack of access to credit by the poor and by small- and medium-sized farms and enterprises is therefore a major deterrent to growth and broad-based development.

Property rights and land rights[33]

Property rights

Property rights refer to the laws and regulations that define rights to ownership, use and transfer of property. A market economy needs secure property rights to eliminate uncertainty and risk associated with undertaking investments relating to the 'property'. It is risky to build a factory, or make improvements on a farm, if these are located on a piece of land whose ownership has not been legally secured. It has been observed in studies covering a broad range of countries that the more secure the property rights, the faster the economic growth. Secure property rights involve *titling*, whereby the owner of a property received a *title*, which is legal evidence of ownership.

Investment and access to credit

An important impact of secure property rights is on investment, which is riskier and therefore less likely to be undertaken when property rights are not secure. For example, in a number of eastern European transition economies (i.e. making a transition from central planning under communism to a market orientation), entrepreneurs who had secured their property rights reinvested 14–40% more of their profits in their businesses than those who had not.[34]

When property rights have been legally secured, there is improved access to credit because the borrower can prove ownership of a property that can be used as collateral (or wealth). In the absence of a legal title to the property, the lender (such as a commercial bank) considers the loan riskier, because of the uncertain ownership of property used as collateral, and will be less likely to offer the loan. The less secure the property rights, the lower the access to credit, the worse the terms (i.e. the higher the rate of interest), and so the fewer the investments financed by credit.

Land rights

Land rights refer to the rights and rules to possess, occupy and use land. Land rights are not necessarily the same as property rights based on titling. In developing countries, land rights are often regulated by *custom* or by *communal ownership*, which do not involve titles. The term *land* here refers to natural resources including the land itself, trees, minerals, pasture, water. Land is an extremely important resource and probably the most important asset (or form of wealth) for the poor. It is also the basis of much of economic activity as it plays a role in agricultural production, provides work for most of the rural population, enables people to get credit, and is a source of livelihood for the billions of the rural population of developing countries.

Yet only 30% of the world's population has legally registered rights to their land and home.[35] At the same time, more than 2.5 billion people depend on community lands for their survival. These lands, covering more than

[32] Consultative Group to Assist the Poor (CGAP) (www.cgap.org).

[33] Land rights are a syllabus term in Chapter 20.

[34] *World Development Report 2005*

[35] https://www.worldbank.org/en/news/opinion/2018/08/16/why-strengthening-land-rights-strengthens-development

50% of the world's surface, are held, used or managed collectively. In Africa, 90% of rural lands have no documentation.[36]

In the 1990s, Hernando de Soto, a Peruvian economist, argued that weak or non-existent property rights were responsible for the failure of developing countries to grow. He therefore recommended the establishment of property rights including land registration and titling as a key strategy for economic development. Reasons why secure property rights would lead to growth included security for investments, better access to credit, development of efficient markets for buying and selling property, increases in tax revenues and increased foreign investment.

This thinking was in line with the market-based supply-side policies that were being advocated at the time, focusing on the development of markets (see Chapter 13). The World Bank together with aid organisations embarked on the process of land titling in many countries, often aimed at privatising land ownership and developing markets for land. Yet subsequent experience showed that whereas these ideas may apply to urban and other properties, there are serious doubts over their applicability to land, particularly agricultural land in developing countries.

Property rights as a western concept with limited applicability to land in developing countries

In 2005, a special commission partly chaired by Hernando de Soto investigated the results of titling. Its conclusions, published in 2008, noted that 'the benefits of titling schemes had been exaggerated and that it may be inappropriate to simply transplant the western concept of property rights into the legal systems of developing countries'.[37]

Some reasons why land titling did not work as expected are the following:

- Titles were often captured by elites who could pay higher prices.
- Once titles are established taxes may be imposed; if the poor cannot pay them they are forced to sell their property or land.
- If the poor borrow using the land as collateral and are unable to pay the loan for example due to a poor harvest they will lose the land to the lender.
- Any temporary economic hardship such as a poor harvest or a drop in commodity prices may force the poor the sell the land.
- Titling increases the market value of land making it less affordable to the poor.

In all these cases the poor find themselves landless. Titling schemes therefore often led to increased inequalities. Instead of increased security for the rural poor they often led to less security and impoverishment.

The commission recognised that the vast majority of the world's poor live in the informal economy, thus not benefitting from a legal order. Moreover, titling programmes neglected forms of tenure based on custom and collective rights which 'could be highly legitimate and effective in guaranteeing security of tenure'.[38] Therefore

> *'To ensure protection and inclusion of the poorest, a broad range of policy measures should be considered. These include formal recognition, adequate representation, and integration of a variety of forms of land tenure such as customary rights, indigenous peoples' rights, group rights, and certificates . . .'*[39]

These rights go far beyond the simple property rights of titling and will be considered in Chapter 20.

The problem of land grabs

For years inadequate or insecure land rights led to the problem of *land grabs* in many developing countries, especially in sub-Saharan Africa, involving the buying or leasing of large pieces of land by individuals, large companies or multinational corporations. This process intensified in 2008–2009 when high food price volatility encouraged the development of large-scale agriculture and land acquisitions for this purpose, and continues to this day. The land that is taken over from a legal point of view is either public or state land or land that is used by custom or communally.

The result of such practices is the risk of eviction and displacement of millions of poor farmers. In addition, it contributes to decreased food security as large farms

[36] https://www.oxfam.org/en/why-indigenous-and-community-land-rights-matter-everyone

[37] https://journals.openedition.org/poldev/2026

[38] https://journals.openedition.org/poldev/2026

[39] CLEP (High-Level Commission on the Legal Empowerment of the Poor) (2008) *Final report*, New York: United Nations.

often produce commodities for export, rather than food products produced by small farmers.

The World Bank (see Chapter 20) has contributed to this process by encouraging the privatisation and takeover of land by large agribusiness firms that continue to dispossess small farmers and undermine food security.[40]

TEST YOUR UNDERSTANDING 19.5

1. Identify the key institutional weaknesses of many developing countries that were discussed in this section and examine their implications for the ability of these countries to grow and develop.
2. Select a developing country of your choice. Examine the extent to which this country is affected you identified in question 1.

Gender inequality

The problem of gender inequalities

Many development countries face serious gender inequalities, or inequalities between women and men, and girls and boys. These relate to control of resources and access to opportunities, and are the result of discrimination against girls and women, with profound consequences for growth and development.

Gender inequalities result in serious deprivations and limited access of women and girls to social, economic and political opportunities compared with men and boys. Eliminating the deprivations and creating conditions of equality of opportunities is called *empowerment*. The empowerment of women is very important in economic development, because not only does it improve the well-being of girls and women affected, but it also gives rise to major external benefits that spill over into other areas influencing growth and development.

Gender inequalities in health and education

In education, girls do not always have the same opportunities as boys, particularly in low-income families where resources are insufficient to educate all children equally. Moreover, cultural factors further dictate that a girl's education is not as important as a boy's, since girls are often not intended to seek work in the labour market on reaching adulthood. In some societies, educated girls are less marriageable.

Table 19.4 provides data on three indicators showing gender inequalities in education: adult literacy, mean years of schooling (for people aged 25 years and older), and expected years of schooling (for a child of school-entrance age). The countries are listed in order of decreasing GNI *per capita* (US$ PPP) (shown in the last column). With the exception of Colombia, in all the countries shown, adult female literacy is lower than adult male literacy. The data on mean years of schooling indicate that in most countries girls receive fewer years of schooling than boys. Colombia is an exception here too, as well as Armenia where the mean years are the same for girls and boys.

Interestingly, the data on expected years of schooling present a different picture for the countries in the upper half of the table, where expected years of schooling is greater for girls than for boys. This suggests that over time, the gender disparities in literacy in these countries will diminish: the girls who are in school today will become part of the literate population in the future. However, this applies to countries with a relatively higher GNI *per capita*, suggesting that in countries with lower incomes gender equality in education will not be realised in the foreseeable future.

Gender inequalities in the labour market

Lower levels of education and skills place women at a great disadvantage relative to men, as they do not have qualifications to take on jobs requiring skills. Even if a woman has the same level of skills as a man, discrimination prevents her from securing a high-level job. Women in low income countries are more likely to work in the informal economy, which means that apart from receiving lower incomes, they cannot receive the benefits offered to workers in the formal sector (job security, legal protection of workers' rights, pensions, etc.) and are subject to the exploitation that often accompanies work in an unregulated labour market. Further, women are far more likely than men to have unpaid responsibilities (child-rearing, household chores,

[40] https://www.oaklandinstitute.org/sites/oaklandinstitute.org/files/highest-bidder-eng-low-res.pdf
https://www.brettonwoodsproject.org/wp-content/uploads/2019/03/The_highest_bidder_takes_it_all-2.pdf

Country	Female adult literacy rate (% of females aged 15 and above) 2015–2016	Male adult literacy rate (% of males aged 15 and above) 2015–2016	Mean years of schooling Female 2017	Mean years of schooling Male 2017	Expected years of schooling Female 2017	Expected years of schooling Male 2017	GNI *per capita* US$ PPP 2017
China	94.5	98.2	7.6	8.3	14.0	13.6	16800
Colombia	94.9	94.4	8.5	8.1	14.9	14.3	14120
Peru	91.2	97.2	8.7	9.7	13.9	13.6	12900
Sri Lanka	91.0	93.0	10.3	11.4	14.1	13.6	12520
Ecuador	93.3	95.4	8.6	8.8	15.4	13.9	11350
Armenia	99.6	99.7	11.7	11.7	13.4	12.6	10060
Morocco	58.8	78.6	4.5	6.5	12.0	12.8	8050
Bolivia	88.6	96.5	8.2	9.7	14.0	14.0	7350
India	60.6	81.3	4.8	8.2	12.9	11.9	6950
Angola	60.7	82.0	-	-	11.0	12.7	6450
Moldova	99.1	99.7	11.5	11.7	11.9	11.4	6100
Zambia	56.0	70.9	6.5	7.4	12.0	13.0	3900
Chad	14.0	31.3	1.2	3.4	6.4	9.5	1920
Uganda	71.5	85.3	4.7	7.2	11.0	12.2	1820
Sierra Leone	37.7	58.7	2.7	4.3	9.3	10.2	1510
Burundi	83.1	88.2	2.7	3.7	9.3	10.2	730

Source: *GNI per capita US$ PPP World Bank, World Development Indicators*
Literacy rates CIA The World Factbook
Mean years of schooling and expected years of schooling United Nations Development Programme, Human Development Report 2018

Table 19.4: Gender inequalities in education

and subsistence farming, i.e. cultivating food for the family, not for sale in the market, which is usually performed by women).

Gender inequalities in inheritance rights and property rights

In many developing countries inheritance rights and property rights are passed mainly to men. Whereas many countries have passed laws ensuring equality between the sexes in inheritance rights, these laws are often not enforced. The same countries usually also restrict the rights of women to own property. Land is overwhelmingly owned by men, even in societies where agricultural work is mainly the responsibility of women.

Gender inequalities in access to credit

Women face serious restrictions in obtaining credit. Over and above the problem of discrimination, lack of property rights means that women have little if anything they can

REAL WORLD FOCUS 19.5

The problem of 'missing women'

Figure 19.6: India. Marathi women perform a folkdance protesting against female infanticide dedicated to the goddess Durga and the empowerment of women and girls

Whereas the ratio of women to men in France, the United Kingdom and the United States is greater than 1.05, in some less developed countries it is much lower: 0.95 in Egypt, 0.94 in Bangladesh, China and West Asia, 0.93 in India and 0.90 in Pakistan. A ratio of women to men of 1.05 means that there are 105 women for every 100 men. The main reason for this is biological: women have higher survival rates than men. A ratio of 0.90 means that there are 90 women for every 100 men. Based on these figures, it is easy to calculate how many more women these countries should have had if they had had the same ratio of women to men as in developed countries. The shortfall in the number of women is known as the problem of 'missing women', i.e. they are 'missing' due to unnecessary deaths. More than two decades ago it was estimated that there are more than 50 million women missing in China alone, and many more than 100 million missing in Asia and North Africa.[41] More recent data estimate 65 million missing women for India alone, where the ratio between women and men has been worsening in several states in recent years.[42]

According to recent forecasts of missing women, these are expected to reach 150 million by 2035. It is expected that newly missing women will be over three million a year every year until 2050.[43]

The most important reason for missing women has been higher female infant and child mortality (deaths) due to less health care and poorer nutrition of girls compared to boys in families where there is not enough income to cover the needs of all the children. Another reason which is increasingly growing in importance arises from prenatal (prior to birth) sex selection which occurs when female foetuses are aborted.

The authors of the study of India's missing women[44] take their argument further, noting that due to the missing women problem, India's voting population is significantly reduced, resulting in a gender bias in policies that favour men.

> *'Since politicians respond to the preferences of the existing electorate, the danger is that electoral competition will ensure that they end up choosing policies in favour of their average voters who happen to be male . . .*
>
> *True political representation of women cannot be ensured unless preferences of women get significant attention. The adverse sex ratio of the Indian electorate makes it impossible for women's welfare to feature in political agendas and policymaking.'*

Applying your skills

Select a country you are interested in and research the problem of missing women. How many women are estimated to be missing? Are the numbers increasing or decreasing over time? Does the government have any policies in place to prevent this from continuing? What policies would you recommend?

[41] Amartya Sen (1999) *Development as Freedom*, Oxford University Press.

[42] https://www.brookings.edu/blog/up-front/2019/04/25/what-indias-65-million-missing-women-mean-for-the-state-of-its-democracy

[43] John Bongaarts and Christophe Guilmoto, 'How many more missing women? Excess female mortality and prenatal sex selection: 1970-2050' *Population and Development Review*, 2015, vol 41, issue 2

[44] Kapoor, M., and Ravi, S. (2013) *Women voters in Indian democracy: A silent revolution.* Available at SSRN 2231026

use as collateral. Moreover, their low earning power makes them far less attractive candidates to receive credit from the point of view of credit institutions (banks).

Gender inequalities in income, wealth and poverty

A number of factors combine to ensure that women's incomes are on average substantially lower than men's: lower levels of education and skills, discrimination against women in the labour market so that women are paid less for the same work, and women's work in the informal sector. Moreover, lack of inheritance and property rights of women ensures that most wealth (such as land and other assets) is concentrated in the hands of men.

These factors contribute to making women among the poorest parts of the population in many developing countries. Poverty among women is particularly pronounced when women are heads of households (i.e. there is no man in the household earning an income); because of their very low earning potential, female-headed households are located in the poorest regions, with limited or no access to sanitation, clean water, health services, etc.

TEST YOUR UNDERSTANDING 19.6

1 Identify some of the key areas of gender inequalities.

2 a Explain some external benefits of women's empowerment.

b Use an externality diagram to show how positive consumption externalities of education and health care for women are a type of market failure.

c Identify government policies that can help correct this market failure.

3 Explain why Amartya Sen elsewhere refers to women as 'active agents of change'.

Inappropriate governance

According to a World Bank report, **governance** is 'the process through which state and nonstate actors interact to design and implement policies within a given set of formal and informal rules that shape and are shaped by power. This report defines *power* as the ability of groups and individuals to make others act in the interest of those groups and individuals to bring about specific outcomes.'[45]

Governance is not about *what* is done for economic growth and development, but rather *how* it is done. It is about the effectiveness of government, but also it involves the relations between government and society, and how they interact to make decisions. According to researchers on this topic, *good* governance consists of six principles:[46]

- **Participation** – the extent to which the stakeholders affected by policies are involved in making decisions and in the implementation of decisions.
- **Fairness** – the extent to which rules apply to everyone in society equally.
- **Decency** – the extent to which the formation and implementation of rules does not harm or humiliate anyone.
- **Accountability** – the extent to which political figures and decision-makers are responsible to society for the actions and their statements.
- **Transparency** – the extent to which decisions made by government are clear and open.
- **Efficiency** – the extent to which scarce resources are used without waste, delays or corruption.

Good governance is important because according to studies making cross-country comparisons, better governance is related to more investment and greater economic growth. The effectiveness of government, the efficiency of bureaucracy and rule of law are positively related to economic performance and adult literacy, and negatively related to infant mortality.[47]

In view of the above, the lack of effective governance in many developing countries is a major barrier to growth and development.

[45] World Bank World Development Report 2017

[46] Goran Hyden, Julius Court and Mease, Kenneth. (2004) 'Making sense of governance: empirical evidence from 16 developing countries', Overseas Development Institute.

[47] Julius Court (2006) 'Governance, development and aid effectiveness: a quick guide to complex relationships', Overseas Development Institute.

Corruption

The World Bank defines corruption as 'the abuse of public office for private gain'. Corruption can take many forms, including bribery, patronage, construction kickbacks, procurement fraud, extortion, false certification, cronyism, nepotism and embezzlement. It occurs everywhere in the world, but is especially pronounced in countries where the legal system, mass media and the system of public administration are weak and underdeveloped, conditions which tend to be more prominent in less developed countries.

The scope of corruption is vast. In 2018, the United Nations estimated that the global cost of corruption is at least US$2.6 trillion per year, corresponding to 5% of global gross domestic product.[48] More than US$1 trillion is spent on bribes alone; this figure includes international flows of bribes, mainly involving multinational corporations.

Corruption around the world is monitored by Transparency International (TI), which is an organisation that measures levels of corruption and transparency. Using public opinion surveys that gauge respondents' views on the degree of corruption in their country, TI compiles an index annually, the Corruption Perceptions Index, which ranks more than 150 countries on the basis of perceived levels of corruption.[49]

Corruption is increasingly recognised as a major barrier to growth and development; it is well documented that higher levels of corruption are found in countries with low *per capita* incomes and low rates of growth. As in the case of political instability, the causality runs in both directions: high levels of corruption are associated with lower investment and growth, which in turn perpetuate the conditions that underlie corruption.

Why is corruption associated with lower growth and poorer development performance? We can distinguish the following factors:

- corruption, in the form of a payment made to obtain something, is like a tax, which in effect makes private investments more costly, thereby reducing the overall level of investment, and lowering the rate of growth
- bribes often have to be paid in order to obtain basic services (such as bribes to teachers and health workers); when this happens they work like a *regressive* tax, because lower income people must pay proportionately more than higher income people in order to obtain the service (i.e. the bribe is a higher fraction of their total income); the result is to deprive the poor of basic social services, thereby going clearly against development objectives
- unlike tax funds, that are paid to the government and that become available for use in socially desirable activities, bribes go into the pockets of public servants and politicians, and are not available to the government for the provision of social services; moreover, corruption that takes the form of bribes for tax evasion further deprives the government sector of resources that would have been available for investments in education, health and infrastructure
- corruption can result in misallocation of resources as government officials may accept bribes to pursue uneconomic projects (such as dams and power plants), while neglecting investments in basic social services (health care facilities, education, sanitation, clean water, etc.)
- corruption prevents the operation of competitive markets, because it restricts the entry of new firms that must pay bribes to begin operations; this is another factor hindering private investment
- corruption reduces hopes for environmental sustainability as government officials accept bribes in order to bypass environmental regulations, authorise the abuse and destruction of natural resources, or finance environmentally unsound projects
- corruption damages the people's trust in the state, its institutions and leadership, and encourages contempt for the rule of law.

[48] https://www.un.org/press/en/2018/sc13493.doc.htm

[49] The interested student can access the Corruption Perceptions Index by going to Transparency International's home page at www.transparency.org/

Unequal political power and status

Political power

Countries differ in the type of political system they have. A political system is a set of legal institutions that define how a government is structured and functions. There is a very broad variety in types of political systems, including monarchies, democracies, republics, oligarchies, and others, with varying forms of legal, constitutional and organisational arrangements.

Whatever the type of political system, what is important is a country's political structure, involving the relationships between various groups within a society and the degree of political power they control. Elite groups within a society, whether these are landowners, industrialists or bankers, may influence the kinds of growth and development policies that can be pursued, and these differ broadly among developing countries. Any group that has disproportionately great political power also has the corresponding ability to influence the government toward making decisions that favour its own interests rather than the interests of the economy or the interests of disempowered, vulnerable excluded groups whose interests are underrepresented.

In general, it is difficult to carry out a development programme without the support of political elites. This poses difficulties for governments undertaking policies that might threaten the interests of the elites.

Social class and status

All societies have some type of *social stratification*, which consists of an arrangement of the population in layers (or strata) on the basis of an unequal distribution of income, wealth, power and prestige. The top layer has the most income, wealth, power and prestige, while the bottom layer has the least. Class and caste systems are two very important ways that social stratification is achieved. An important difference between class and caste is the degree of social mobility, or the possibility of moving from one social layer to another.

In a *caste system*, an individual's social position is determined by birth, and cannot be changed. The most well-known caste system is that in India, where discrimination based on caste still persists though it is illegal. Caste systems may be based on religion (such as Hinduism in the case of India's caste system) or on race (such as skin colour in cast systems imposed in Africa by European colonisers, where white skin was the superior caste), or on ethnic factors.

In a *class system*, by contrast, an individual's social position is determined by factors over which there is some control (education, income, type of work). A caste system is clearly more harmful to growth and development, as it predetermines an individual's lot in life, however rigid class systems can have similar impacts. Both are associated with the exercise of prejudice and discrimination, with the upper classes or castes discriminating against the lower ones. Moreover, in caste or rigid and highly hierarchical class systems, the top caste or class often uses its power to prevent the lower castes or classes from attaining social mobility and moving out of their disadvantaged social and economic positions.

Racial and ethnic issues

Racial and ethnic ties form very strong bonds between the people who are members of particular racial or ethnic social groups. *Race* refers to people who have been grouped together on the basis of certain physical characteristics (such as skin colour). *Ethnicity* refers to people who share common cultural characteristics (such as language, dress, food, values, place of origin), and who share a common cultural identity. Sometimes race and ethnicity coincide (for example, African-Americans and Japanese-Americans). Race and ethnicity sometimes lead to the exercise of prejudice and discrimination as well as intolerance, and at times result in social tensions and violent conflicts.

Political stability

Political stability refers to stable government and its ability to withstand forcible removal from power. Countries differ enormously with respect to their degree of political stability. The presence of political stability is associated with higher rates of growth and improved development outcomes for the following reasons.[50]

[50] Although political stability is not easily quantifiable, it is routinely monitored for countries around the world by international organisations (such as the World Bank) by use of several indicators that measure people's perceptions of the likelihood that the government in power could be overthrown or destabilised by violent or unconstitutional means. Examples of some of these indicators include civil war, revolutions, assassinations, internal conflicts, ethnic tensions, demonstrations, riots, strikes, frequency of elections, government crises, major constitutional changes, cabinet changes, religious tensions, and others. The lower the likelihood that the government could be brought down, the greater the perceived political stability.

- A stable government is necessary for effective government decision-making and for implementing economic and other policies that have continuity over some years, creating a stable economic environment.
- Political instability creates an environment of uncertainty related to economic policy, property rights, possibility of expropriation, and taxation rules, all of which make both domestic and foreign investments far riskier, thereby reducing investments.
- Political instability often leads to an outflow of financial capital as people seek safety for their financial assets (capital flight), depriving the country of its scarce financial resources and contributing to balance of payments deficits.
- Political instability increases vulnerability to hunger and famine, as it deprives governments of the capacity to provide relief, while resources are diverted to military or police activities. The 1984–1985 famine in Ethiopia resulting in over one million deaths (with another seven million people severely affected) was caused as much by political instability, internal war and violence as by drought. According to the World Bank, the most important cause of famine in developing countries is not poverty or low agricultural output, but military conflict.

There is a close relationship between political instability and levels of income: in general, low levels of income *per capita* are associated with higher levels of political instability. The causality (cause and effect) runs in both directions: political instability is a cause of low incomes because it gives rise to low economic growth; and low incomes are a cause of political instability because they lead to widespread dissatisfaction and frustration with economic conditions for which the government is held responsible.

TEST YOUR UNDERSTANDING 19.7

1 Explain why each of the following is a barrier to developing countries' efforts to grow and develop:

   a gender inequalities

   b inappropriate governance

   c corruption

   d political power and status

   e (optional) political instability.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Some consider the One Laptop per Child non-profit initiative an example of 'appropriate technology' though the project has its critics. Research the OLPC. Do you believe it is an example of appropriate technology? Why or why not?

2 Research and find a developing country that has experienced an improvement in its distribution of income and another country where this has worsened. Identify the factors responsible for the increased and decreased equality in distribution.

3 Section *Low levels of human capital* lists many barriers to education and good health. Research a country of your choice and identify barriers that country faces. Find facts and statistics to support the existence of the barriers you identify. Share your findings with a classmate and listen to their research too.

4 Research a developing country of your choice that faces gender inequalities with regard to a variety of factors (schooling, health, property ownership, etc.). For that particular country, suggest reasons why these inequalities exist and assess the current trends and possibilities in the future for women to close the inequality gap.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

> Chapter 20

Strategies to promote economic growth and economic development

BEFORE YOU START

- Recalling what you have learned in microeconomics, macroeconomics and international economics, can you think of measures and policies governments can use to encourage economic growth and economic development in their country?
- Would some measures and policies be more appropriate for developing countries, and others more appropriate for developed countries?

This chapter critically examines a broad variety of policies and strategies that developing countries may pursue in order to promote their growth and development. Wherever relevant, reference is made to particular Sustainable Development Goals (SDGs) that relate to the policy in question.

Photo: Zanzibar, Tanzania. Children splashing in sea water

20.1 International trade strategies

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and evaluate trade strategies including: (AO3)
    - import substitution
    - export promotion
    - economic integration
    - trade liberalisation

Import substitution

Import substitution, also known as *import substituting industrialisation*, is a growth and trade strategy where a country begins to manufacture simple consumer goods for the domestic market to promote its domestic industry (for example, shoes, textiles, beverages, electrical appliances, etc.). Import substitution depends on protective measures (tariffs, quotas, etc.) preventing the entry of imports that compete with domestic producers.

Many Latin American countries adopted these policies from the 1930s onward. By the 1950s and 1960s, most developing countries around the world were pursuing industrialisation based on import substitution. It was attractive because most currently developed countries had used import-substituting policies in the initial phases of their industrialisation. The theoretical justification was provided by the infant industry argument recommending the use of trade barriers to protect 'infant' domestic firms against competition from imports (see Chapter 15).

Import-substitution policies and consequences

Import substitution policies had the following common characteristics and consequences:

- **High levels of protection of domestic firms, inefficiency and resource misallocation.** Protection took mainly the form of tariffs, quotas and import licences. The resulting lack of competition led to high costs and inefficiency in private and public sector industries, resource misallocation and high prices for consumer goods.
- **Overvalued exchange rates.** Many countries overvalued their exchange rates (set them at a higher level than the free market level) reducing the price of imports and increasing the price of exports (see Chapter 16). The objective was to allow firms to import capital inputs more cheaply; however, it had two negative effects:
    - Cheap capital imports led to capital-intensive production methods (inappropriate technologies), unemployment and growth of the informal economy in urban areas.
    - It made agricultural exports more expensive, worsening rural poverty.
- **Too much government intervention in the economy.** Most import-substituting countries relied heavily on industrial policies (interventionist supply-side policies; see Chapter 13), with strong intervention in the form of protective trade barriers, overvalued exchange rates, subsidised credit, tax allowances, production subsidies, wage subsidies, price controls, etc., as well as extensive public ownership of firms and industries (fertilisers, steel, petrochemicals, cement, banking and financial services, infrastructure, and many others). This led to serious resource misallocation and inefficiencies in production.
- **Neglect of agriculture.** Agriculture was neglected, and due to the failure to make agricultural investments, there was an increased need for food imports.
- **Deterioration in the balance of payments.** The balance of payments deteriorated because of:
    - increasing imports of capital equipment as inputs in production
    - increased need for food imports
    - outward flow of financial capital due to profit repatriation of foreign multinational corporations (profits taken to the home country).
- **Encouragement of capital-intensive production methods.** It was believed that this would result in more rapid growth. There was no effort to provide support for small entrepreneurs more likely to use labour-intensive techniques.

- **Negative impacts on employment and income distribution.** Capital-intensive technologies and the neglect of small producers increased unemployment and contributed to the development and growth of the informal economy, along with worsening income distribution and increasing poverty.
- **Limited possibilities for growth over the longer term.** In spite of some growth in the early periods, there came a point when it was no longer possible to grow through import substitution. This was due to serious inefficiencies (high costs of production). Many firms enjoying protection never 'grew up' to become efficient, low-cost producers, firms that should have closed down were kept going, while others that should have been set up or expanded were not.

By the 1970s and 1980s, there was general agreement among economists that import substitution had not lived up to expectations. It therefore began to be abandoned in favour of *export promotion*.

Today import substitution is not practised as a general strategy for growth and industrialisation, however there may be *selective import substitution*, that may include industrial policies for particular infant industries.

Export promotion

Export promotion is a growth and trade strategy where a country attempts to achieve economic growth by expanding exports. Like import substitution, export promotion was based on strong government intervention, justified by the idea that this is necessary to help countries develop a strong manufacturing sector oriented towards exports.

The experience of export promotion

Export promotion strategies evolved gradually as an extension of import substitution. In many cases, the industries that became the strongest exporters were the ones that had received strong import-substituting protection.

The economies that first turned to export promotion included China, Indonesia, Japan, Malaysia, Singapore, South Korea, Thailand and others; they are known as the *Asian Tigers*. While each economy was unique in the blend of policies, some typical policies included the following:

- **Financial assistance to targeted key industries,** including:
  - ***Targeting of export industries*** that used increasingly higher skill and technological levels.
  - ***Industrial policies to support export industries,*** including investment grants, production subsidies, exemptions from tariffs of imported inputs, tax exemptions, export subsidies, and special benefits for export oriented multinational corporations.
  - ***Provision of incentives to the private sector for R&D in high technology products*** to encourage the development of domestic skills and technologies appropriate to local conditions.
- **Strong government intervention in the economy**, including:
  - ***State ownership and control of financial institutions (banking and insurance)*** to provide subsidised credit to the industries being promoted, such as lower interest rates and other favourable borrowing terms.
  - ***Large public investments in key areas*** including education and skills, R&D, and expansion and modernisation of transport and communications infrastructure.
- **Requirements imposed on multinational corporations,** intended to maximise the benefits of foreign direct investments, such as the promotion of R&D, transfer of desired and targeted technologies into the domestic economy, training of domestic workers, and the use of local inputs where possible.
- **Exchange rate management**, involving *undervalued currencies* that encourage exports while making imports more expensive.

These policies resulted in immensely successful export performance and achievement of very high growth rates, due to increases in aggregate demand. Since the 1950s, the Asian Tigers have been the fastest growing economies in the developing world. In addition, they succeeded in making significant improvements in their levels of economic and human development.

Factors behind the success of export promotion over import substitution

Why were the Asian Tigers so successful?

- **Expansion into foreign markets**, taking advantage of economies of scale.
- **Emphasis on diversification.** Beginning with support for simple, labour-intensive goods (for example,

textiles and clothing), industrial policies later supported diversification based on increasing skill and technology levels (see below)

- **Major investments in human capital**, including education, training and skills.
- **Appropriate technologies.** Governments supported R&D for the development of appropriate technologies, as well as transfer from abroad of technologies appropriate to local conditions.
- **Increased employment**, resulting from the use of labour-intensive technologies.
- **No balance of payments problems**, due to significant increases in exports and export earnings.

Possible disadvantages of export promotion

- Exporting countries may become overly dependent on exports so that in the event of recession in the major trading partners, exports will fall leading to a drop in aggregate demand with consequent recession.
- There may be efforts to maintain low wages to keep labour costs low thus making exports more competitive; hence workers may not benefit from growth that exports make possible.
- Strong exports over a long period of time lead to a trade surplus corresponding to trade deficits in trading partners, possibly leading to trade protection by the trading partners who feel threatened by excessive imports.

Economic integration

SDG 17.10 (goal 17, target 10) states, 'Promote a universal, rules-based, open, non-discriminatory and equitable multilateral trading system under the World Trade Organization . . .'.

The growth of preferential trade agreements

This topic was introduced in Chapter 15. In the last few decades, there has been a very large increase in the number of bilateral and regional trade agreements around the world. The number of trade agreements reported to the World Trade Organization (WTO) grew from 20 in 1990 to 159 in 2007, and 270 by 2019. One reason is that many countries are becoming frustrated with what they believe is the slow progress made by the WTO (see Chapter 15). Another is that developing countries see in trading blocs the possibilities of enjoying the benefits of free trade, bypassing obstacles created by rich country trade protection, while maintaining some of the benefits of trade protection (toward non-members).

Chapter 15 offered an evaluation of trading blocs in terms of their advantages and disadvantages as a method to achieve free trade. (You should refer to this discussion as it is closely related to our present topic). We will now evaluate trading blocs as a strategy to achieve growth and development. To do so, we must make the distinction between regional and bilateral trade agreements.

Regional free trade agreements (FTAs): potential benefits for growth and development

Economists generally agree that free trade agreements have the greatest potential to help developing countries achieve growth and development when they involve:

- *regional* agreements
- geographical closeness
- similar level of development and technological capabilities
- similar market sizes
- shared commitment to co-operation.

These conditions allow countries to achieve the benefits of integration we studied in Chapter 15. Regional groupings allow countries to expand their markets (achieving economies of scale) and to diversify production and exports. Larger markets increase domestic and foreign direct investment. When countries are at a similar level of development with similar technological capabilities as well as similar market sizes, the new competition created by increased imports is more 'fair' and easier to deal with (it does not involve 'unfair' competitive advantages of foreign firms caused by lower costs due to use of more advanced technologies, greater managerial know-how, larger size due to larger home markets).

If there is a shared commitment to co-operation, there are several policies that can be pursued jointly by members so they can further benefit from their integration. They can invest in transport infrastructure needed for trade, as well as in energy and water supplies needed for growth and development. They can collaborate on R&D projects and new technology development that would be mutually beneficial. They can work together on environmental issues of common interest.

These factors greatly increase the likelihood that integration will lead to increased growth and more development. While it is difficult for all these conditions to be met in practice, it is not surprising that we usually find neighbouring countries forming regional blocs such as in Latin America (MERCOSUR), southeast Asia (ASEAN), eastern and southern Africa (COMESA), central Africa (CEMAC), Central America (CAIS), etc.

Bilateral free trade agreements (FTAs): risks for growth and development

Most of the trade agreements in existence are bilateral, and most bilateral agreements are between developing and developed countries that are *not* usually in the same geographical region (though there are exceptions). The developed countries mainly involved are the United States, which has agreements with a number of developing countries, the European Union (which acts as a unit) that also has agreements with developing countries and transition economies, and Japan, with agreements mainly in the Asia-Pacific region.

A bilateral agreement has the potential to provide a developing country with access to the developed country market, and the prospect of gaining such access is why developing countries enter into such agreements. However, this potential comes with risks:

- The developing country must make equal and matching cuts in tariff and other barriers, often much greater than those required by WTO agreements. This puts even efficient developing country firms at a competitive disadvantage because they are forced to compete with a lower cost developed country. The result may be to destroy even efficient local firms.
- When many developing countries form FTAs with the same developed country to gain market access, the advantage each hopes to gain individually is lost, as they must now all compete with each other for the developed country market. Thus, increases in exports may be limited.
- Increased imports and only slightly increasing exports may result in trade deficits, balance of payments problems and increasing foreign debt. They may also result in greater unemployment, worsening income distribution and increased poverty.
- Bilateral negotiations put developing countries at a disadvantage due to weaker bargaining power compared to the multilateral negotiations of the WTO where they can join together and present their interests as one.
- The developing country must agree to other requirements that may not be in its best interests (such as freer rules on foreign direct investment, stricter rules on copyright and patent laws).
- Bilateral agreements divide developing countries by creating different interests. They also weaken regional trade agreements when a member country makes a bilateral agreement with a third country.

According to the United Nations Conference on Trade and Development (UNCTAD),[1] developing countries are better off pursuing regional trade agreements. The trend toward bilateral trade agreements:

> *'threatens the viability of existing regional cooperation arrangements among developing countries, and, most importantly, the options available to these countries for pursuing their national development strategies.*
>
> *FTAs can result in some export gains, and possibly increased FDI (foreign direct investment) flows, but the size and durability of these benefits is highly uncertain, as are the net gains for trade and output growth. This is because the FTA will most likely lead to an increase in imports, with implications for the trade balance and, in some cases, the external debt position. Moreover, if future … FTAs are modelled on those that have been negotiated so far, it is likely that they will considerably reduce or fully remove policy options and instruments available to a developing country to pursue its development objectives.'* [2]

[1] UNCTAD is a United Nations organisation concerned with international trade issues in developing countries.

[2] UNCTAD, *Trade and Development Report, 2007* (available at http://www.unctad.org/en/docs/tdr2007_en.pdf).

Trade liberalisation

A fourth important trade strategy is trade liberalisation. This will be presented below in connection with market-based policies.

TEST YOUR UNDERSTANDING 20.1

1. a. Define import substitution and explain why it has an inward orientation.
   b. What were some factors that led most developing countries to adopt import substitution as an industrialisation strategy in the 1950s?
   c. Explain some of the key policies that were associated with import-substituting strategies.
   d. Evaluate the effectiveness of these strategies with respect to their impacts on economic performance and economic growth and development.
2. a. Define export promotion and explain why it has an outward orientation.
   b. What were some of the countries that adopted an export orientation during the 1960s, and what prompted them to do so?
   c. Explain some of the key policies that were associated with export promotion.
   d. Evaluate the effectiveness of these policies with respect to economic performance, export growth and economic growth and development.
3. Discuss the potential advantages and disadvantages of
   a. regional trade agreements, and
   b. bilateral trade agreements as a strategy to promote growth and development.

20.2 Diversification and social enterprise

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and evaluate the strategies of (AO3)
  - diversification
  - social enterprise

Diversification of economic activity

Diversification involves a reallocation of resources into new activities that broaden the range of goods or services produced. As a country grows and develops, the relative share of the primary sector in GDP usually shrinks, becoming progressively replaced by manufacturing (industry) and later by services. The decline in the relative share of the primary sector's contribution to GDP is made possible by the *diversification* of production into manufacturing as well as services (see Chapter 3 for an explanation in terms of the income elasticity of demand.) As the country diversifies into manufacturing and services, its exports are likely to become more varied accordingly. Such diversification of production and exports has been taking place particularly in higher income developing countries.

The importance of expanding into higher value-added production

The concept of 'value added' refers to the value of a good that is added in each step of a production process.[3] For example, suppose a country produces cocoa beans. The beans are cleaned, roasted, and shelled, then ground into a paste, called cocoa liquor. A portion of the liquor goes into hydraulic presses that remove the cocoa butter, leaving behind a paste that is ground into a fine powder for use in the baking industry. The cocoa butter is mixed with the other portion of the liquor together with sugar,

[3] For a numerical example, see footnote 4, Chapter 8

vanilla and milk, all of which are kneaded well and placed into machines called conches that stir and shake the mixture under heat for some days. When the mixture cools, it is converted into chocolate bars or is used in baking and confectionary industries.

Each step in this process *adds value* to the cocoa beans. If a country that produces cocoa beans goes on to produce chocolate, it adds value to its primary commodity and *diversifies* into chocolate manufacturing.

Adding value in diversification is important because it means:

- engaging in more varied production activities
- creating employment opportunities
- establishing new firms involved with manufactured goods
- expanding into activities requiring higher skill and technology levels.

If a country produces primary commodities and exports these in their 'raw' form, it misses the benefits listed above. It becomes trapped in methods and kinds of production that keep it in a state of stagnation, with limited prospects for change. Developing countries must therefore seek out new opportunities for processing of raw materials and manufacturing.

The issue of too much specialisation and the importance of diversification is related to a key criticism of the theory of comparative advantage (HL only; see Chapter 14). If a country has a comparative advantage in the production of primary products, specialisation and trade according to comparative advantage would seriously limit its growth and development prospects.

Overspecialisation versus diversification

In Chapter 19 (Section 19.2), we saw that overspecialisation in a narrow range of primary commodities is a major barrier to growth and development. As Table 19.1 illustrates, a number of the countries that are overspecialised are rich in natural resources. Ironically this resource wealth often works against them. It has been observed that since the 1960s *resource-poor developing countries have been growing faster than resource-rich countries*. This surprising trend has been termed the 'curse of natural resources' because it suggests that resource-rich countries might have been better off without these natural resources. Examples of countries that have experienced low and sometimes negative rates of growth in spite of their abundant resource supplies include Russia (rich in oil, natural gas, metals and timber), Nigeria, Mexico and Venezuela (rich in oil), Congo and Sierra Leone (rich in mineral deposits), South Africa (rich in oil and mineral deposits), and others.

The reasons behind the better performance of resource-poor countries can be found in their earlier diversification into manufacturing. The inability to rely on production and export of primary commodities made them turn early on toward labour-intensive manufacturing, together with investments in human capital and appropriate technologies. Resource-rich countries, on the other hand, became heavily dependent on primary commodities, with short-term volatility of export revenues, the need to resort to external borrowing and accumulation of large debts, and balance of payments difficulties.

The benefits of diversification

It is hardly possible to overemphasise the importance of diversification as a strategy for growth and development. It permits countries to achieve the following important objectives:

- **Sustained increases in exports.** Increase in exports must be maintained over long periods. This can only be achieved through diversification into markets with a sustained increase in global demand, which commodity exports do not satisfy.
- **Development of technological capabilities and skills.** Diversification provides incentives to acquire new technologies and higher training, education and skill levels, which are very important for growth and development. This was one factor behind the spectacular success of the Asian Tigers.
- **Reduced vulnerability to short-term price volatility.** Diversification protects countries against losses from fluctuating export prices.
- **Use of domestic primary commodities.** Countries that already produce primary products are in a special position to use these as the basis for their diversification into manufacturing, as the domestic availability of the necessary raw materials can work to stimulate industry. Countries that got a major boost to their growth and development through this type of diversification include Malaysia, Thailand, Indonesia, China, Chile and Mauritius.

Social enterprise

A **social enterprise** is a type of commercial organisation that aims to achieve particular social goals to improve people's well-being and promote social change. Social enterprises may be either for-profit or not-for-profit organisations. If they are for profit their primary objective is to achieve their social goals, not to maximise profit. They try to be commercially viable (cover all their costs) rather than rely on grants or donations, by selling the services or products they provide. If they make profits these are put back into the enterprise rather than received as profit income by the owners.

Social enterprises exist everywhere in the world but are becoming increasingly popular in developing countries where they focus on anti-poverty programmes and other efforts to meet important social, economic or environmental goals.

Social enterprises operate in a broad variety of areas, including education, health, social care, agriculture, fisheries, forestry, energy, clean technology and transport. Most of the workers are relatively young and there is a high female participation rate.

Provision of microfinance (see below) is considered to be a type of social enterprise. Muhammad Yunus, the founder of microfinance, used the expression *social enterprise* in connection with his work.

TEST YOUR UNDERSTANDING 20.2

1 Outline the benefits that a country can expect from expanding into higher value-added production.

2 Explain how diversification can help a country grow and develop.

3 Explain how social enterprise can contribute to economic development.

REAL WORLD FOCUS 20.1

Social enterprise: Bambike

Bambike produces hand-made bicycles made of bamboo in the Philippines. Their website states

> *'Bambike is a socio-ecological enterprise based in the Philippines that hand makes bamboo bicycles with fair trade labor and sustainable building practices. Our bamboo bike builders aka Bambuilders come from Gawad Kalinga, a Philippine based community development organisation for the poor, working to bring an end to poverty. We have programmes that include scholarships, sponsoring a pre-school teacher and a weekly feeding programme for children, as well as a bamboo nursery for reforestation. Bambike is a company that is interested in helping out people and the planet, dedicated to social and environmental stewardship. Our goal is to do better business and to make the greenest bikes on the planet.'*

Figure 20.1: Manilla, Philippines. A bambike

In addition, Bambike produces sunglasses, tumblers and cups all made of bamboo. They also offer guided tours of historic neighborhoods and other destinations on bamboo bicycles.

Source: *www.bambike.com*

Applying your skills

1 While Bambike is a for-profit organisation it has important social goals. Use the information above and on its website to outline Bambike's social goals.

2 Research and identify one or more other social enterprises in a country of your choice. Discuss how the activities of the social enterprises help achieve the country's development goals.

20.3 Market-based policies

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO3)
- explain and evaluate market-based policies including (AO3)
    - trade liberalisation
    - privatisation
    - deregulation

Trade liberalisation, privatisation, deregulation and more

The spectacular successes of the Asian Tigers made a sharp contrast to the performance of most developing countries. In the early 1980s, many were showing poor export and growth performance, and were highly indebted. This was the time when monetarist/new classical economists were introducing market-based supply-side ideas, emphasising the importance of limited government intervention and the competitive free market.

As a trade and growth strategy, limited government intervention meant **trade liberalisation** (elimination of trade barriers to achieve free trade) and a free market approach in the domestic economy. This free market approach to growth and development came to be known as the *Washington Consensus*, because it was shared by the World Bank, the International Monetary Fund, the United States Congress, and a number of US agencies (all based in Washington, DC). The main policies recommended by the Washington Consensus included:

- **trade liberalisation** involving moving to free trade by lowering and eliminating tariff and other barriers to trade
- **privatisation of state enterprises,** such as transport, oil, gas
- **deregulation** of labour and product markets; this meant adopting market-based supply-side policies for labour (reducing labour union power and unemployment benefits, reducing or eliminating minimum wages; see Chapter 13); it also meant removing barriers to enter product markets.

Additional policies included:

- encouraging floating exchange rates; no exchange rate management
- reduced restrictions to foreign direct investments by multinational corporations
- limiting borrowing by the government; keeping budget deficits under control
- restricting the use of industrial policies.

These measures involved freeing up markets and cutting back on the role of government. Countries should no longer follow the highly successful export promotion policies based on strong government intervention. They were based on the idea that reliance on market forces and free trade improves efficiency and the domestic and global allocation of resources, and increases economic growth.

Since the 1980s, many developing countries increasingly adopted liberalising policies by following these policy prescriptions. Examples include Argentina, Brazil, China, Chile, India, Kenya, Sri Lanka, Tanzania, Turkey, the countries of East Asia, and many more. These countries did not completely abandon their interventionist policies, but instead began a gradual reduction of government intervention, with some countries liberalising more, or more rapidly, than others.

The effects of economic and trade liberalisation

By the 1990s there was evidence that liberalisation of trade and the economy was not bringing about the expected benefits:

Limited benefits for export growth and diversification

Many countries found themselves losing export shares in world markets (the proportion of exports in relation to global exports). The losses were greatest in Africa. The UNDP notes that if, by the early 2000s, Africa still had its 1980 share of world exports, its exports would be greater by US$119 billion (in constant $ in terms of the year 2000); this is equivalent to about five times the amount of aid provided by donors in 2002.[4]

On the whole liberalisation policies did not help developing countries diversify production into increased manufacturing. In 2000, just five developing countries were responsible for two-thirds of developing country low technology manufactured exports, while only six were responsible for more than four-fifths of developing country medium and high technology manufactured exports.[5] In most Latin American and African countries, growth of manufacturing exports was slow to moderate, with no significant change indicating diversification into manufacturing. In some Latin American countries there was *a decline* in the relative share of manufacturing. The countries that fared best were those that had already developed significant export sectors (the East Asian countries).[6]

Partly, these negative effects of liberalisation were due to the trade protection policies of developed countries on developing country exports, including protection of agriculture and tariff escalation (see Chapter 19). They were also due to growing reliance on free market policies. Remember, the great successes of the East Asian countries were based on industrial policies involving strong government intervention. With less government support, many developing countries were not able to perform well.

Limited effects on economic growth

According to well-known development economists, 'There is little evidence that open trade policies – in the form of lower tariff and non-tariff barriers to trading – are significantly associated with economic growth.'[7] Furthermore, 'Perhaps the most comprehensive assessment of the links between economic growth and trade liberalisation undertaken to date concluded that there is no clear link between them. This means that the projected benefits are merely hypothetical.'[8]

According to the United Nations Development Programme:

> *'One of the prevailing myths of globalisation is that increased trade has been the catalyst for a new era of convergence . . . However, successful integration is the exception rather than the rule, and trade drives global inequality as well as prosperity. For the majority of countries the globalisation story is one of divergence and marginalisation.'*[9]

Increasing income inequalities and poverty within developing countries

There is clear evidence that economic and trade liberalisation resulted in greater income inequalities and poverty. A World Bank study noted that trade liberalisation leads to lower income growth among the poorest 40% of the population, but higher income growth for the higher income groups. In other words, it helps the rich get richer and the poor get poorer.[10]

The reason is that economic and trade liberalisation creates both 'winners' and 'losers'. When new export markets are opened up, those who find employment in the production of export goods will be better off; people who find jobs in a growing formal sector will also gain; people with some education and skills may also gain as they are better able to exploit new opportunities in the more competitive environment. However, there will also be those who will become worse off. They include:

- less educated or illiterate people, who are unable to compete in the new environment
- poor people who lack collateral (wealth), and who cannot get credit to open or expand a business to take advantage of new opportunities
- people in remote geographical areas with no transport links to markets
- people who have nothing to export, and no possibilities of producing for export

[4] United Nations Development Programme, *Human Development Report 2002* (available at http://hdr.undp.org/en/reports/).

[5] United Nations Development Programme, *Human Development Report 2005* (available at http://hdr.undp.org/en/reports/).

[6] S. M. Shafaeddin (2005) 'Trade liberalisation and economic reform in developing countries: structural change or de-industrialisation?', Discussion Paper, UN Conference on Trade and Development (UNCTAD), (available at www.unctad.org).

[7] F. Rodriguez and D. Rodrik (1999) 'Trade policy and economic growth: a skeptic's guide to the cross-national evidence', Discussion Paper, National Bureau of Economic Research.

[8] L. Alan Winters (2000) 'Trade liberalisation and poverty', Paper, Centre for Economic Policy Research, London, and Centre for Economic Performance, London School of Economics.

[9] United Nations Development Programme, *Human Development Report 2005 (available at http://hdr.undp.org/en/reports/).*

[10] M. Lundberg and L. Squire (1999) *Inequality and Growth: Lessons for Policy*, The World Bank.

- people in agriculture who switch to producing commodities for export, making themselves more vulnerable to wide fluctuations (volatility) in commodity prices
- people who lose their jobs as public employees due to cutbacks in the public sector (in Zimbabwe, they are referred to as the 'new poor')
- people who lose their jobs in the private sector as firms close when they are unable to compete with imports of products produced by large firms in richer countries
- people who become unemployed due to privatisation of public enterprises, which fire workers to lower costs
- people affected by cuts in government spending on merit goods, forced by a greater reliance on market forces
- people affected by lower levels of social protection caused by supply-side policies (such as lower minimum wages, lower protection against being fired, etc.; see Chapter 13)
- people forced into the informal economy, where wages are lower and social protection is non-existent, due to removal of trade protection leading to closure of formal sector firms that can no longer compete (in Zambia, for example, formal employment fell by 15% in the decade of the 1990s[11]).

International trade theory recognises that free trade is likely to give rise to both winners and losers. However, it argues that since the overall gains will be greater than the overall losses, the gainers can compensate the losers, with the result that no one need be worse off. Yet, in the real world, such compensation rarely (if ever) takes place.

The free market approach of the Washington Consensus was questioned even by some individuals within the World Bank. Joseph Stiglitz, as Chief Economist of the World Bank, writes the following:

> *'The neoliberal model*[12] *accords the government a minimal role, essentially one of ensuring macroeconomic stability, with an emphasis on price stability, while getting out of the way to allow trade liberalisation, privatisation, and getting the prices right. Many of these policies are necessary for markets to work well and contribute to economic success, but they are far from sufficient. Some aspects of the neoliberal model might not even be necessary conditions for strong growth, and if undertaken without accompanying measures . . . they may not bring many gains and could even lead to setbacks. Some countries have closely followed the dictates of the neoliberal model, but have not seen especially strong economic performance. Other countries have ignored many of the dictates . . . and have experienced among the highest rates of sustained growth the world has ever seen.'*[13]

A new consensus: trade and market liberalisation with government intervention

Since the late 1990s, supporters and critics of market liberalisation have been moving towards a new consensus according to which there should be a mix of markets with government intervention to support growth and development. The following are some of the ideas in this view:[14]

- Governments must support education, health services and infrastructure development, as well as research and development (R&D) and transfer of technology for both industry and agriculture.
- Large budget deficits should be avoided, but if contractionary fiscal policy is needed, it should not affect spending on education, health and infrastructure.
- Governments must pay attention to the effects of policies on income distribution, and must pursue policies that promote income equality and alleviation of poverty.
- Governments must provide a proper regulatory framework for markets to work effectively; for example, there should be effective regulation for competition (otherwise privatisations may lead to the development of private monopolies).

[11] Winters (2000) 'Trade liberalisation and poverty'.

[12] By the term 'neoliberal model', Stiglitz is referring to the free market approach of the Washington Consensus.

[13] Joseph E. Stiglitz (1998) 'Knowledge for development: economic science, economic policy, and economic advice', Annual World Bank Conference on Development Economics, Washington, DC, April 1998.

[14] Joseph E. Stiglitz (1998) 'More instruments and broader goals: moving towards the Post-Washington Consensus', Annual Lecture, World Institute for Development Economics Research, Helsinki, 1998; and 'Towards a new paradigm for development strategies and processes', Prebisch Lecture, UNCTAD, Geneva, 1998.

THEORY OF KNOWLEDGE 20.1

Moral issues of trade liberalisation in developing countries

In the Theory of knowledge 15.1 (Chapter 15), we considered the moral judgement that is implied in the recommendation that countries adopt free trade. Here, our topic is broader and perhaps more serious, and involves the moral implications of the trade (and economic) liberalisation policies recommended for developing countries (and sometimes forced upon them; see below) by developed ones.

The issues are numerous, but can be divided into two broad categories: trade protection policies of more developed countries that prevent developing country access to their markets (agricultural protection of farmers, high tariff barriers, tariff escalation) and the trade liberalisation policies of the Washington Consensus, the WTO and bilateral trade agreements.

Nobel Prize-winning economist Joseph Stiglitz, referring to the free market approach to international trade that since the 1990s has dominated development policies, writes the following about moral and ethical issues in relations between developed and developing countries:

> *'Economists have long bought into the importance of self-interest not only in explaining behaviour, but also in yielding efficient outcomes. But economists have also long been aware of the limitations of these perspectives. Not only does the self-interest/market paradigm often fail to generate efficient outcomes, but even when it does, these outcomes may not [be consistent with] with notions of social justice …*
>
> *Ethics in the relationship between developed and less developed countries dictates that the developed countries treat the less developed countries fairly, aware of their disadvantaged economic position, and acknowledging that taking advantage of one's own economic power inevitably will hurt the poor within developing countries. [There are] several instances where, in global economic relationships, this precept has been grossly violated: an international trade agenda set to advance the interests of the more developed countries, at least partially at the expense of the less developed – so much so that on average the world's poorest region was actually worse off at the end of the last round of trade negotiations;[15] and an international environmental agreement that provided that those rich countries who today are polluting more be entitled to continue polluting more into the future.'*[16]

Thinking points

- Do developed country societies have a moral obligation to help developing ones (especially the poorer ones)?
- Consider the following question posed by Joseph Stiglitz. 'At one level, it is natural for a country to pursue its own interests. But . . . at what point does this pursuit of a country's own interest (or, as is more frequently the case, special interests within one's country) at the expense of the poor, become a moral issue?'[17]
- Are developed countries morally justified in promoting bilateral free trade agreements with developing countries when they refuse to give up protection of their farmers?
- How fair are the trade rules of the WTO?
- Do the organisations of the Washington Consensus (the World Bank, IMF and US government) bear any moral responsibility toward developing countries for mistaken policies that in some cases were damaging to the poor of those countries (such as countries in sub-Saharan Africa)?

[15] This is a reference to sub-Saharan Africa.

[16] Joseph Stigltiz, (2005) 'Ethics, economics advice and economic policy', Initiative for Policy Dialogue, 24 October.

[17] Joseph Stiglitz (2005) 'Ethics, economics advice and economic policy'.

- Efforts must be made to promote institutions such as property and land rights, an effective tax system, and effective banking and credit system (see below).
- Developed countries must assist economic development by increasing foreign aid and providing increased access to their markets for developing country exports.
- Developing countries should receive special treatment by international trade agreements under the World Trade Organization regarding removal of rich country trade protection measures (for example, in agriculture).
- According to the new consensus, government intervention is important to help create the conditions for markets and trade to work to the advantage of developing countries.

TEST YOUR UNDERSTANDING 20.3

1 a Define trade liberalisation, referring to its objectives.

  b Outline the main ideas behind the market-based policies recommended for developing countries.

  c Explain the connections between the market-based supply-side policies discussed in Chapter 12 and the Washington Consensus.

2 Discuss the effects of liberalising policies on export growth, diversification, economic growth and income distribution.

20.4 Interventionist policies: redistribution and provision of merit goods

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and evaluate interventionist policies including (AO3)
  - redistribution policies including
    - tax policies
    - transfer payments
    - minimum wages
- explain and evaluate provision of merit goods including
  - education programmes
  - health programmes
  - infrastructure (energy, transport, telecommunications, clean water and sanitation)

Redistribution policies

SDG 10 states 'Reduce inequality within and among countries'. This is the objective of redistribution policies.

Tax policies

SDG 17.1 (goal 17, target 1) states 'strengthen domestic resource mobilisation . . . to improve domestic capacity for tax and other revenue collection'.

In Chapter 19 we saw that developing countries often face a number of difficulties regarding taxation, leading to low levels of revenues, inequities, and negative effects on resource allocation. In view of these difficulties it is important that developing countries undertake reforms that will improve their taxation systems.

The International Monetary Fund recommends that developing country governments should:[18]

- increase the progressivity of the tax system, which is generally lower in comparison with developed countries
- gradually expand the coverage of personal income taxes
- expand the use of indirect taxes on luxury goods and goods that create negative externalities, such as cigarettes, alcohol and energy based on fossil fuels
- increase taxation from capital income (profits), which is essential to ensure progressivity
- impose or increase taxes on real estate and land
- take measures to reduce tax evasion.

Transfer payments

Transfer payments were discussed in Chapter 12 where we saw that they are an important method used to improve income distribution and reduce poverty.

SDG 1.3 (goal 1, target 3) states 'Implement nationally appropriate social protection systems and measures for all, including floors, and by 2030 achieve substantial coverage of the poor and the vulnerable.'

Building on this, the World Bank and International Labour Organization (ILO)[19] called on countries around the world to design and implement by 2030 *universal social protection*. Universal social protection means access by an entire population to social protection which includes child benefits, pensions for older persons, and benefits for maternity, disability work injury or unemployment.

In 2019, four billion people or more than half of the world's population had no access to even one social protection benefit. 45% of the global population had access to one benefit, one-third of children had a family and child benefit, and only 28% of people with disabilities had a benefit. Old age pensions were the most prevalent, with 68% of older persons receiving pensions.[20]

The benefits take mainly the form of transfer payments which are cash transfers or benefits in kind. They are very important because they:

- reduce poverty and encourage social inclusion
- empower individuals and encourage them to make decisions that reflect their preferences rather than the preference of the government or aid or development organisations
- increase incomes that are used to increase demand for locally produced goods, thus also helping local producers
- promote economic growth by raising incomes, consumption and investment that increases aggregate demand
- help provide safety against sudden hardships or emergencies
- improve access to health care and education
- encourage empowerment of women, and delayed marriage
- reduce malnutrition and child mortality
- help bring the poor into the formal economy
- result in reductions of child labour
- build political stability and reduce social tensions
- according to numerous studies, do not reduce the incentive of adults to work.[21]

A special policy that has become increasingly popular involves *conditional cash transfers* (CCTs), involving money paid on condition that the households receiving the money undertake activities related to education and health care, often for children. (See Real world focus 6.1 in Chapter 6 on the Bolsa Familia programme in Brazil.) *Non-conditional cash transfers* do not impose restrictions, providing flexibility to households to manage their expenditures freely in accordance with their needs. At the time of writing there did not appear to be a consensus on which of the two is more successful as an antipoverty policy as there is a large variety of programmes of both types around the world whose results have not yet been fully evaluated.

Universal social protection systems are costly and may be beyond the means of many developing countries suggesting that they cannot be fully implemented over the short term. Moreover, their full effectiveness depends on government policies to address the crucially important issues of providing schools, hospitals,

[18] https://www.imf.org/en/Publications/FM/Issues/2017/10/05/fiscal-monitor-october-2017

[19] The ILO is an agency of the United Nations.

[20] https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_669390/lang--en/index.htm

[21] https://www.brookings.edu/blog/future-development/2018/12/19/from-one-to-many-cash-transfer-debates-in-ending-extreme-poverty
https://www.worldbank.org/content/dam/Worldbank/Event/safetynets/1.%20Lindert_UCTs%20and%20CCTs%20for%20SSNCC.pdf
https://www.usp2030.org/gimi/USP2030.action

and infrastructure including roads, clean water and sanitation, trained doctors and nurses as well as good training for doctors and nurses, all of which are needed for development. Further issues involve difficulties in the design and implementation so that the money will reach the population groups that are most in need.

Minimum wages

Minimum wages were introduced in Chapter 4. According to standard theory, whereas they are designed to support incomes of unskilled workers, they give rise to unemployment. In line with this thinking, in the 1980s and 1990s market-based supply-side policies emphasised reducing or eliminating minimum wages to increase employment (Chapter 13). However since then, a growing number of studies have indicated that in practice such job losses do not occur unless minimum wages are set at very high levels. As a result, since the late 1990s, many countries around the world introduced or increased their minimum wages.

In view of this change in thinking about minimum wages, countries wanting to reduce income inequalities see minimum wages as policy that can help. In many countries the question is not whether or not to have minimum wages but how to best design a minimum wage system. Addressing this issue of policy design the International Policy Centre for Inclusive Growth (IPC-IG)[22] notes the following:

- Minimum wages should be set by governments after consulting with representatives of workers and employers in order to take all relevant points of view into account.
- In deciding on the level of the minimum wage it is important to consider the needs of workers and their families; to monitor and evaluate the effects; and to have a mechanism allowing possible changes every year or two in order to make necessary adjustments.
- It is important to establish measures to ensure compliance and enforcement, to avoid work at wages below the legal minimum.

Provision of merit goods

Education and health services have been discussed at length in Chapters 6, 12, 18 and 19. They are merit goods whose importance for growth and development cannot be overemphasised. Education and health appear in the following SDGs:

- SDG 1.A.2 (goal 1, target A, indicator 2) states 'Proportion of total government spending on essential services (education, health, social protection)'. This is a measure of resource mobilisation to end poverty.
- SDG 4 states 'Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all'.
- SDG 3 states 'Ensure healthy lives and promote well-being for all at all ages'.

Education and health services as merit goods

Positive externalities of education for growth and development

Education provides many benefits for society, making the social benefits of education far greater than individual benefits; external benefits arise from positive consumption externalities:

- Economic growth is partly a benefit of education, as the benefits of education extend to society in the form of increased labour productivity and greater output.
- Education contributes to improvements in the quality of physical capital (i.e. technological advances), because knowledge can be applied to research and development, and especially to the development of technologies appropriate to local economic, ecological and climate conditions.
- Education results in lower unemployment, lower absenteeism from work and increased international competitiveness; and it attracts foreign direct investment.
- Education leads to increased political stability, an important condition for economic growth and development.
- Education provides further social benefits, such as a lower crime rate and a better quality of life.
- The education of women in particular promotes their increased participation in the labour force, lower birth rates (fewer children), leading to lower rates of population growth and reduction of poverty.

[22] The IPC-IG is a partnership between the United Nations and the Government of Brazil to promote social policies for developing countries. https://www.ipc-undp.org/pub/eng/PIF42_Minimum_wage_global_challenges_and_perspectives.pdf

- Some benefits of education spill over into health:
    - The education of women leads to improvements in maternal health and reductions in maternal mortality (deaths).
    - The education of mothers results in healthier children through improved health care and better nutrition and lower child mortality (deaths)
    - Schools teach children basic principles of hygiene and sanitation, which improve the levels of children's and family health.

The importance of elementary education and universal literacy

Many studies have shown that East Asian countries (including China, Indonesia, South Korea, Thailand, and others) have invested more heavily in education than other countries with comparable income levels. It is widely believed that this has been a key factor behind their superior growth and development performance.

One of the most important investments in education involves achieving universal literacy rates. The highly successful East Asian countries began their initial drive for growth and development by pursuing high literacy rates, and later gradually increasing investments in secondary and tertiary (university) education in line with developments in their technological capabilities.

According to estimates made by the World Bank, in the case of countries at a relatively low level of development, the greatest contribution to economic growth can be made by primary (elementary) school education. Yet developing countries often invest too heavily in higher (secondary and university) education, while ignoring the basic elementary level. This misallocation of resources results in:

- an *international brain drain*, whereby university graduates from developing countries migrate to developed ones, resulting in loss of human capital. In effect, poor country resources build human capital in rich countries.
- an *internal brain drain*, whereby
    - highly educated individuals who cannot find employment in their field work in unrelated areas requiring lower skill levels (for example, a doctor or engineer working as a taxi driver)
    - doctors and medical personnel educated in government-funded institutions do not work in public medical institutions intended to provide free services to lower income groups, but instead work in the private sector which serves wealthy patients
    - highly educated individuals apply their skills and abilities to research and technology development in areas that are more relevant to the needs of developed countries because these are more prestigious, while ignoring local technological needs (such as building low-cost schools, hospitals and housing, etc.).

Both give rise to a significant misallocation and waste of scarce public resources.

Positive externalities of health care for growth and development

An improved level of health also provides benefits beyond the individual ones:

- Health leads to greater worker productivity and therefore greater output and economic growth.
- Healthy people do not transmit diseases, thus lowering the risk of spreading diseases to the community.
- Immunisation benefits not only the immunised person but also the community by lowering the risk of contracting a disease.
- Healthier people provide more benefits to the community through more active and productive participation.
- Some benefits of higher levels of health spill over into education, with:
    - increased levels of health and good nutrition improve school attendance and performance in school, leading to longer time spent in school
    - healthier individuals make better use of the knowledge and skills they possess
    - better health means a longer lifespan, and so a longer time during which the benefits of education can affect the economy and society.

Appropriate intervention in education and health care in developing countries involves mainly direct government provision of services that are free of charge. An estimated 100 million people are forced into extreme poverty by having to pay for health care, while many die unnecessarily because they cannot afford to pay for it.[23] In addition to health services, also important are investments in sanitation, clean water and sewerage,

[23] https://www.who.int/news-room/detail/13-12-2017-world-bank-and-who-half-the-world-lacks-access-to-essential-health-services-100-million-still-pushed-into-extreme-poverty-because-of-health-expenses

as well as legislation making education compulsory up to a certain age. Advertising and persuasion as well as nudges (HL concept) may also help convince parents of the benefits of education and preventive health services (such as immunisation). Note that legislation, advertising and persuasion would be entirely useless in a situation where there are no schools or health clinics; this is why governments must step in and provide these services directly in areas where they do not exist.

According to Gro Harlem Brundtland (former Director General of the World Health Organization):

> *'In many countries, while those with money are able to access good healthcare and education, hundreds of millions of ordinary people are denied life-saving health services or are plunged into poverty because they are forced to pay unaffordable fees for their care. The burden is particularly felt by women and children, who have high needs for services but the least access to financial resources. In some countries, poor women and their babies are even imprisoned in hospitals because they can't pay their medical bills after giving birth.*
>
> *The solution to this problem is simple: universal public services provided free at the point of delivery. Unfortunately, powerful political interests often oppose this proven way to reduce inequalities.*
>
> *Overcoming this opposition and launching equitable public services requires a large investment of public financing and political capital by governments and political leaders. As well as improving social indicators, accelerating economic growth and reducing inequalities, this is also a smart political choice that can strengthen social cohesion and provide an enduring legacy.'*

According to Oxfam, public services of education and health care must be:

- universal, meaning that everyone should have access to them
- free at the point of use, meaning that users of the services should not have to pay for them
- public, not private, meaning they should be provided by the government
- able to prioritise services important to women, and promote women as workers (see below on women's empowerment)
- accountable to those they serve.[24]

Infrastructure

Infrastructure is addressed in SDG 6: 'Ensure availability and sustainable management of water and sanitation for all' and SDG 9: 'Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation'.

Infrastructure and economic growth and development

Infrastructure increases productivity (output per worker) and lowers costs of production. Good road and railway systems save time and effort in transporting goods and services, allowing more output to be transported and production costs to be lowered. The availability of effective telecommunications permits faster and easier communications, enabling economic activities to be carried out more efficiently. Irrigation contributes to higher yields (output per unit of land) and expansion of agricultural output.

The availability of infrastructure also facilitates modernisation and diversification of the economy. The growth of electronic communication and data exchange has contributed to more efficient practices and expansion of manufacturing, financial services and government economic activities. The availability of power (electricity) allows for increases in worker productivity through the introduction of simple electrically powered machines and equipment. Safe water sources, sanitation and sewerage permit countries to diversify into the production of processed foods. The quantity and quality of infrastructure are important for a country's international competitiveness, because they determine shipping costs. The availability of good quality infrastructure also attracts foreign direct investment.

Infrastructure provides services that are essential for maintaining a basic standard of living. Safe water supplies, sanitation and sewerage systems have major effects on levels of health of a population, contributing to the reduction of avoidable illnesses and premature deaths.

Transport services also affect health and education by bringing people in remote rural areas closer to educational and health facilities. Transport increases employment opportunities by allowing the movement of people across longer distances, thus increasing incomes and contributing to the alleviation of poverty. It also facilitates access to markets, reducing the time and costs of transporting goods. The availability of transport can

[24] https://www.oxfam.org.nz/sites/default/files/reports/Public%20Good%20or%20Private%20Wealth%20-%20Oxfam%202019%20-%20Summary.pdf

be crucially important to integrating people in remote and isolated rural areas into the market economy.

Availability of water supplies, along with infrastructure supplying energy (electricity and gas), have major impacts on gender equity. When these services are not available, women and girls are forced to spend a large proportion of their time carrying water and fuel-wood; in some African countries these activities take as much as two-thirds of women's household time. The availability of piped water supplies, and electricity and gas, by freeing time, increases school enrolment among girls. In the case of women, the availability of these services leads to increased employment outside the home, and reduced fertility (fewer children in the family) and hence reduced poverty.

In addition, the availability of safe energy sources (electricity and gas) results in less indoor air pollution (arising from the burning of polluting fuels for cooking and light), with strong positive effects on the health of women and children, who spend more time indoors. The introduction of irrigation over large areas, by increasing yields (output per unit of land) similarly increases incomes and contributes to raising people out of poverty. Construction and maintenance of infrastructure contributes to creating employment opportunities for the people who work to construct and maintain the facilities, thus also increasing incomes. The employment-creating effects are especially strong in the case of labour-intensive methods used to build roads.

TEST YOUR UNDERSTANDING 20.4

1 Discuss how
   - a tax policies,
   - b transfer payments, and
   - c minimum wages can be used to redistribute income.

2 Examine how the contributions of
   - a education programmes,
   - b health programmes, and
   - c provision of infrastructure can contribute to economic growth and development.

20.5 Foreign direct investment and multinational corporations (MNCs)

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and evaluate foreign direct investment (AO3)

Foreign direct investment (FDI) is investment by firms based in one country (the home country) in productive activities in another country (the host country) with control of at least 10 per cent of the firm in the host country. A firm that undertakes foreign direct investment is referred to as a **multinational corporation (MNC)**, because it operates in more than one country. A 'corporation' is a type of firm composed of a legal entity that is separate from the individuals who own it.

The scope and growth of multinational corporations

Multinational corporations run business operations in both the home country and in other (host) countries. Historically, MNCs have been active since about the middle of the 19th century. Their importance grew in the 1950s when US multinationals stepped up their investments in Europe as part of European post-war reconstruction. In the last several decades their growth has been explosive. In the early 1990s, there were an estimated 37 000 multinational corporations globally; by 2009, they had increased to 82 000 and employed 80 million people in their foreign affiliates alone.[25]

For most of the 20th century, FDI originated in developed countries and was also directed mainly towards developed countries. Since the 1980s, developing countries have been receiving an increasing share of inflows, approaching half of the total annual FDI inflows as Table 20.1 indicates.

[25] According to the OECD more recent data are not available (https://www.oecd.org/industry/ind/MNEs-in-the-global-economy-policy-note.pdf).

Host region	% of world	% of developing countries
World	100.0	
Developed countries	49.9	
Developing countries	46.9	100.0
Africa	3.0	6.4
Asia	33.3	70.9
Latin America and the Caribbean	10.6	22.6
Transition economies*	3.3	

*Include South East Europe, the Commonwealth of Independent States, Georgia

Source: *UNCTAD, World Investment Report 2018*

Table 20.1: Geographical distribution of foreign direct investment inflows, 2017

The larger of the multinational corporations are enormous in size. In 2017, the top ten non-financial multinational corporations (ranked by size of revenues) had revenues of more than US$2.7 trillion.[26]

Multinational corporations are estimated to produce 33% of global output, nearly two-thirds of which is in the home country and the remaining third in their foreign affiliates. They account for more than half of world exports and nearly half of imports.[27]

Yet FDI remains a small share of total private investment in developing countries; total investment by local firms tends to be far greater than total investment by multinational corporations. This raises an interesting question. If FDI forms only a small share of total private investment in developing countries, why is it the subject of heated discussions and controversy? The answer is that FDI is qualitatively very different from local investment, because of the very large size of MNCs, their significant economic and political power, and their superior technical and managerial expertise, know-how and technologies.

Foreign direct investment is the most important source of foreign finance flows to developing countries as a whole. However, for many low-income developing countries that are almost completely bypassed by MNCs, foreign aid is the main source of foreign finance.

Why MNCs expand into economically less developed countries

Multinational corporations expand into developing countries (as elsewhere) in the hope of securing higher profits. Developing countries offer possibilities for MNCs to:

- **Increase sales and revenues.** Some developing countries have large or rapidly growing markets (for example, China, India and countries in Latin America), which offer the potential for large increases in sales and revenues.
- **Bypass trade barriers.** Producing in countries with trade barriers allows MNCs to bypass these and secure access to local markets.
- **Lower costs of production.** Labour costs take up a large proportion of total production costs, and developing countries generally have lower labour costs than in developed countries. This is a key reason for example, why the United States has multinational corporations operating in Mexico.
- **Use locally produced raw materials.** If an MNC needs raw materials in the form of natural resources for its production, it is far less costly to obtain them locally than to import them, on account of transportation costs.
- **Further their activities in natural resource extraction.** Some MNCs specialise in the extraction of natural resources (oil, aluminium, bauxite, etc.). Many developing countries are very rich in natural resources (for example in Africa), and therefore it is natural for MNCs to want to locate in such resource-rich countries.

Developing country characteristics that attract multinational corporations

Multinational corporations are highly selective in their choice of hosts. They prefer to invest in countries providing them the freedom to pursue their economic interests with the least amount of government

[26] Finance Online https://financesonline.com/top-10-richest-companies-in-the-world-in-2018-by-revenue The top ten non-financial firms were Walmart, State Grid Corporation of China, Sinopec Group, China National Petroleum Corporation, Toyota Motor, Volkswagen, Royal Dutch Shell, Berkshire Hathway, Apple Inc, ExxonMobil.

[27] OECD Multinational Enterprises in the Global economy, May 2018 https://www.oecd.org/industry/ind/MNEs-in-the-global-economy-policy-note.pdf

interference, in a safe economic and political environment that minimises uncertainties and potential risks of losses on their investments. They look for:

- political stability and a stable political environment
- a stable macroeconomic environment: low inflation, stable currency, acceptable levels of foreign debt, absence of major balance of payments problems
- favourable tax rules (to ensure low tax payments)
- weak labour protection laws (to lower the cost of labour)
- a liberalised (free market) economy and trade policy with an emphasis on exports
- large markets
- rapid economic growth and expectations of continued rapid growth
- well-functioning infrastructure, including transportation and communications, to facilitate imports and exports
- a well-educated labour force.

It is easy to see that the rapid growth of FDI around the world in the past several decades has been driven by the liberalisation of the global economy and domestic economies of many countries. Since the 1980s, as developing countries turned more and more toward the market MNCs have found it profitable to establish affiliates in hospitable foreign countries that accommodate their goals.

Advantages and disadvantages of FDI for economically less developed countries

Multinational corporations are profit-seeking entities; they are not organisations concerned with the growth and development problems of developing countries. Why then do developing countries view them as a mechanism that can help accelerate growth and development?

Potential advantages of MNCs for host developing countries

- **MNCs can supplement insufficient foreign exchange earnings.** Investment funds flowing into a country from abroad appear as credits in the financial account, and can help offset a current account deficit. As the activities of multinational corporations are usually export oriented, increased export earnings positively affect the current account.
- **MNCs can supplement and improve upon local technical skills, management skills and technology.** When multinational corporations set up affiliates, they bring with them technical and managerial expertise, as well as new production technologies, which can be learned and adopted by the local labour force (workers and managers) and local businesses.
- **MNCs can supplement insufficient domestic savings and increase investment and new capital formation.** The inflows of FDI funds into a country can add to insufficient domestic savings, increasing the amount of investment.
- **MNCs can lead to greater tax revenues in the host country.** If multinational corporations are taxed by the government of the host country, there will be increased tax revenues.
- **MNCs can help promote local industry.** When MNCs buy locally produced inputs, they promote the development of local industries. This may lead to the growth of existing local firms, or the establishment of new local firms.
- **MNCs can increase local employment and help lower unemployment in the host country.** MNCs can increase employment by hiring local workers.
- **MNCs can lead to higher economic growth in the host country.** Increased levels of investment, improved technology and increases in human capital as well as the promotion of local industry and greater tax revenues, can lead to higher economic growth in the host country with increased possibilities for pursuing development objectives.

Potential disadvantages of MNCs for host developing countries

Why the benefits listed above might not come about

- **MNCs may not always supplement insufficient foreign exchange earnings**. While MNCs usually bring foreign exchange into the host country, they also engage in activities resulting in foreign exchange outflows. These include repatriation of profits (profits sent back to the host country); or MNC imports of raw materials and other inputs; or because they finance their activities by borrowing from the parent corporation in the home country, so they must repay the loan plus pay interest. Therefore the net inflows of foreign exchange (inflows minus outflows) may be small.

- **MNCs may not improve on local technical skills, management skills and technology.** The reason is that the links between MNC activities and the local economy are often limited, in which case local workers do not have the opportunity to learn from the MNC. Also, MNCs often hire personnel from the home country.
- **MNCs may not lead to greater tax revenues in the host country.** MNCs enjoy many tax privileges and benefits, often lowering the amount of tax paid. Tax benefits are offered as an incentive to attract MNCs into the host country. Another reason involves the practice of *transfer pricing*, which works in the following way. Many MNCs buy inputs from their various affiliates in other countries. By claiming that the prices they paid to buy inputs is higher than the actual price paid, their profits appear lower. Since the tax paid is a percentage of profit, lower-stated profits mean lower taxes (sometimes significantly lower). It is estimated that lost tax revenues due to transfer pricing are in the billions of dollars each year.
- **MNCs may not help promote local industry.** The operation of MNCs sometimes forces local competing firms to go out of business, or alternatively does not permit new local firms to establish themselves in industries that are directly competitive with the MNC.
- **MNCs may not help lower unemployment in the host country.** If, as noted above, MNCs prevent the development of local industry, then their job-creating impact will be limited. In addition, some MNCs may sometimes import into the host country capital-intensive technologies that are inappropriate to local conditions, thus contributing to unemployment and the growth of the informal economy. (However, some MNCs engage in labour-intensive activities that make extensive use of cheap local labour).

Further possible negative effects of MNCs

- **MNCs and environmental degradation.** MNCs often pursue activities that cause serious environmental degradation. Preferring to invest in countries with few environmental restrictions, they are known to engage in activities that have caused tremendous environmental damage. One of the greatest disasters caused by MNCs involved an explosion in a Union Carbide plant in India in 1984 that killed more than 20 000 people and left more than 100 000 with serious and permanent health problems. While destruction on such a scale is unusual, there are numerous well-documented cases of MNCs undertaking environmentally unsustainable activities. Moreover, MNCs are responsible for the production of the bulk of industrial pollutants (such as chlorofluorocarbons, a main cause of ozone depletion, as well as pesticides, plastics, petroleum, industrial chemicals, and many others). It has been estimated that since 1988, 100 MNCs have been responsible for more than 71% of greenhouse gas emissions.[28]
- **MNCs promote inappropriate consumption patterns in developing countries.** Critics charge that MNCs, through advertising, create new consumption needs and promote inappropriate consumption patterns. This charge applies to the role of MNCs in developed countries as well, but what makes it more powerful in the case of developing countries is that populations plagued by hunger, malnutrition, disease and lack of basic services can less afford to spend their small incomes on unnecessary goods while their basic needs remain unsatisfied. Examples include consumption of soft drinks, sweets, fast foods, white bread, expensive brand name goods, and many others.
- **MNCs may use government resources to build infrastructure needed by MNCs rather than for poverty alleviation.** MNCs sometimes require infrastructure (road systems, ports, telecommunications, etc.) which the developing country must make available if it is to become attractive as a host country. To build these types of infrastructure, it may have to shift some of its scarce resources away from needed merit goods (clean water, sanitation, schools and health care services) and toward infrastructure for MNCs.
- **MNCs may use their economic and political power to bring about policies that may work against economic development.** The very large size of many MNCs gives them exceptional economic and political power that they can use to influence host governments to pursue policies that are in their own interests but against economic development. For example, MNCs are interested in investing in countries that have weak labour protection laws, since this results in lower costs of production; and they are interested in investing in countries with weak environmental regulations, as this allows them to avoid costs associated with environmental protection. When the interests of MNCs and

[28] https://fortune.com/2017/07/10/climate-change-green-house-gases

those of developing countries conflict, developing country governments find themselves in a weak bargaining position because if they do not give in to MNC demands, they will lose the investment to another developing country that is more willing to compromise. For example, in Peru, a mining company pressured the government not to undertake health tests for children living close to the mining operations.

- **Competition between developing countries to host MNCs and the 'race to the bottom'.** Many developing countries compete with each other over which will create better conditions to attract MNCs. Yet MNC demands may conflict with what is in a country's best interests. This has been termed 'the race to the bottom', because the desire to host MNCs may involve sacrifices in terms of needed policies for growth and development.

TEST YOUR UNDERSTANDING 20.5

1 a Describe foreign direct investment (FDI) and multinational corporations (MNCs).

  b Explain why MNCs have an interest in expanding into developing countries.

  c In view of their relatively small share in total private investment in developing countries, explain why MNCs are a highly controversial topic.

2 a Outline characteristics of developing countries that MNCs look for when deciding where to invest.

  b Suggest why low-income countries receive negligible amounts of foreign direct investment.

  c Outline what factors account for the massive growth in foreign direct investment in developing countries in recent years.

3 Discuss advantages and disadvantages of MNCs in developing countries.

4 Outline why some observers refer to the competition between developing countries to attract MNCs as the 'race to the bottom'.

20.6 Foreign aid

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- distinguish between humanitarian aid and development aid (AO2)
- explain and evaluate foreign aid in the form of (AO3)
  - Official Development Assistance (ODA)
  - non-governmental organisations (NGOs)
  - debt relief

SDG 17.2 (goal 17, target 2) states: 'Developed countries to implement fully their official development assistance commitments, including the commitment by many developed countries to achieve the target of 0.7 per cent of ODA/GNI to developing countries'.

Foreign aid is defined as the transfer of funds or goods and services to developing countries with the main objective to bring about improvements in their economic, social or political conditions. Figure 20.2 providing an overview of foreign aid, shows that for such transfers to be considered as foreign aid, they must be:

- *concessional*, which means that the transfers involve more favourable conditions than could be achieved in the market. If the aid involves loans, interest rates are lower and repayment periods are longer than borrowers would get in commercial banks. Also, the aid may involve *grants*, which are gifts of either money or goods and services that do not need to be repaid.
- They must be *non-commercial*, meaning that they must not involve buying and selling (commerce) or other activities concerned with making a profit.

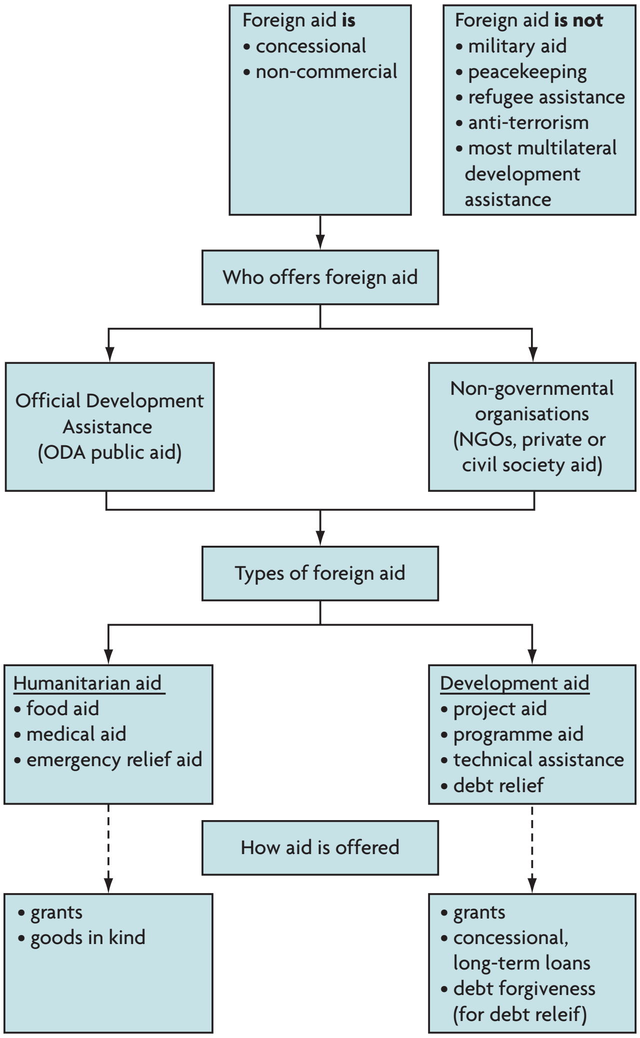


Figure 20.2: Overview of foreign aid

Figure 20.2 also shows what does *not* account as foreign aid, for example military aid, peacekeeping, refugee assistance, and others. These activities are not included under 'foreign aid' because they are not directly concerned with bringing about improvements in the economic, social or political conditions of developing countries.

Under 'Who offers foreign aid' in Figure 20.2, we see there are two sources of aid. The first is *Official Development Assistance (ODA)*, provided by developed country governments, and the second is aid provided by *non-governmental organisations (NGOs)*. All providers of aid are referred to as 'donors' of aid; the developing countries that receive the aid are 'recipients'.

Under 'Types of foreign aid' in the figure, we see that there are two main categories of aid: humanitarian and development.

Humanitarian aid

Humanitarian aid involves aid extended in regions where there are emergencies caused by violent conflicts or natural disasters such as floods, earthquakes and tsunamis. They are intended to save lives, to ensure access to basic necessities such as food, water, shelter and health care, and to provide assistance with reconstruction work in order to help displaced people cope.

Humanitarian aid is extended by donors through grants (sending money as a gift) or through goods-in-kind (food, medical supplies, blankets, etc.).

Development aid

Development aid is intended to help developing countries achieve their economic growth and development objectives. It may involve financial support for specific projects, such as building schools, clinics, hospitals, irrigation systems or other agricultural infrastructure; financial support to sectors, such as education, health care, agriculture, energy, the environment, or others; *technical assistance* in the form of technical advice by specialists such as doctors, teachers, agronomists, or others; as well as debt relief (to be discussed below).

As Figure 20.2 indicates, development aid is extended by donors through grants or through concessional long-term loans as well as debt forgiveness.

Humanitarian aid and development aid are offered by both ODA and NGOs. Whenever it involves financial inflows, these enter as credits in the balance of payments, thus bringing in foreign exchange and helping countries offset possible deficits in their trade balance.

Official Development Assistance (ODA)

Official Development Assistance (ODA), all of which is public in the sense that it comes from government funds, forms the largest part of foreign aid. Most ODA funds (nearly three-quarters) take the form of grants.

ODA funds reach developing countries in three ways:

- through *bilateral aid*, which is the most important way – funds go directly from the donor government to the developing country recipient; examples of bilateral aid agencies are USAID (US Agency for Internal Development) in the United States and DFID (Department for International Development) in the United Kingdom

- through *multilateral aid*, going indirectly from donor governments to international organisations, which transfer the funds to developing country governments
- through NGOs – donor governments transfer ODA funds to NGOs which spend them in developing countries.

The donor countries include most of the members of the Organization for Economic Co-operation and Development (OECD), some members of the Organization of the Petroleum Exporting Countries (OPEC), and more recently also some eastern European countries.

International organisations providing ODA include United Nations agencies; single-issue funds like the Global Fund to Fight Aids, Tuberculosis and Malaria; the International Development Association (IDA, which is an organisation of the World Bank[29]), regional development banks (such as the European Bank for Reconstruction and Development (EBRD), the Inter-American Development Bank (IDB); the International Monetary Fund (IMF) assistance for debt relief under debt relief initiatives.

Donor motives for providing ODA

Donor countries are motivated to provide aid through ODA for a variety of reasons:

- **Political and strategic motives.** For example, during the Cold War, the United States provided aid to restrict the spread of communism. The Soviet Union provided aid to communist states as well as some non-communist states with communist leanings. European powers provided aid to their former colonies. Often aid has been used to support regimes in developing countries that are considered to be 'friendly' to the interests of the donor governments.
- **Economic motives.** Developed countries often regard it to be in their interest to assist countries with which they have strong economic ties. For example, much of Japan's aid is directed towards neighbouring countries with which it has strong trade and investment links. The practice of *tied aid* (to be discussed below), is an important example of economic motives of donors.
- **Humanitarian and moral motives.** Some aid is provided on humanitarian grounds for short-term emergency assistance, such as in the case of famines, wars or natural disasters. Concern about the extent of poverty in developing countries is a motive for allocating aid funds for long-term development purposes.

As a result of these priorities of donors, aid funds do not go to the countries that need them the most. In fact, *middle income developing countries receive ten times as much aid per person in extreme poverty as lower income countries*.[30]

Private (civil society) aid: non-governmental organisations (NGOs)

Non-governmental organisations (NGOs) are the second type of aid flowing into developing countries. Like ODA, they involve concessional flows, but they are all grants (there are no loans that must be repaid).

The World Bank defines NGOs as 'private organisations that pursue activities to relieve suffering, promote the interests of the poor, protect the environment, provide basic social services, or undertake community development'.[31] NGOs are 'private' in the sense that they are not part of any governmental structure; they are not private in the sense of being part of the market system. NGOs are an expression of civil society, and as such are often referred to as comprising a third sector (the first and second being the government and market sectors).

NGOs include a wide variety of organisations, such as charitable organisations, non-profit organisations, nationally based groups with a national or international reach, locally based community groups, or grassroots organisations, and they may operate in developed or developing countries (or both). Among the better-known international NGOs (INGOs) are

29 The International Development Association (IDA) is part of the World Bank but it offers the poorest of developing countries concessional loans, i.e. soft loans – unlike World Bank loans, which are extended on commercial terms.

30 https://www.odi.org/sites/odi.org.uk/files/resource-documents/12392.pdf

31 The World Bank, Operational Directive 14.70, 28 August 1989.

Amnesty International, Greenpeace, Oxfam, Save the Children and World Wide Fund for Nature, also known as the World Wildlife Fund (both abbreviated as WWF).

NGOs in developing countries have grown massively in numbers and in involvement since the 1980s. It is estimated that developing countries now have several tens of thousands of *national* NGOs, and several hundreds of thousands of *community-based* NGOs. A growing number of these now have consultative status with United Nations agencies; from 41 in 1948, the number of NGOs with consultative status today is in the thousands. Most of these are small local groups pursuing development objectives within a relatively small community.

NGOs obtain their funds from private voluntary contributions including private sector corporations and, increasingly, from bilateral and multilateral ODA funds. In other words, more and more of ODA funds are channelled through NGOs, particularly in the case of humanitarian assistance. The reason for this is that NGOs can perform functions that are not performed as effectively by national governments.

NGOs are involved in a vast range of activities, including provision of humanitarian aid in times of crisis, promotion of sustainable development, promotion of community development, service delivery, poverty alleviation, protection of child health, promotion of women's rights, promotion of small-scale entrepreneurs, support of the poor in the informal sector, provision of technical assistance to small farmers, provision of credit to poor people (microfinance), research activities, political advocacy, support for people's movements, and more.

Evaluating foreign aid

Arguments in favour of Official Development Assistance (ODA)

Aid and the poverty cycle

To emerge from a poverty cycle (Chapter 19), poor people and poor communities need the government to intervene by undertaking the necessary investments in physical, human and natural capital. If the government does not have enough tax revenues, the only way the country, or community within a country, can escape the poverty cycle is through foreign aid that makes up for the lack of savings. Very poor developing countries do not have enough funds to make the necessary investments in health care, education and basic infrastructure to help people escape the poverty cycle. Therefore, countries can escape the poverty cycle if foreign aid provides the missing funds for these investments.

Aid and provision of basic services

Even if a country is not caught in a poverty cycle, aid can make resources available for investments in health, education and infrastructure, which can help poor people improve their employment opportunities and improve their incomes. In a number of sub-Saharan African countries, foreign aid is an important component of social budgets. Many of these programmes contribute to significantly limiting the incidence of preventable diseases and reducing infant and child deaths.

Aid and improved income distribution

By focusing on the most disadvantaged groups in society, aid can help improve the relative income positions of the beneficiaries and contribute to improved income distribution.

Aid and economic growth

There is strong evidence that aid leads to economic growth, because it makes possible increased investment and consumption levels, leading to increased volumes of output.

Aid and the Sustainable Development Goals (SDGs)

The provision of aid is crucially important to the achievement of the Sustainable Development Goals (SDGs). Much of aid is closely linked to the achievement of these goals. According to the United Nations Development Programme, it will not be possible for developing countries to achieve the SDGs without enough aid.

Aid, the debt trap and debt relief

Countries that are heavily indebted (have high levels of debt) face serious negative consequences for their growth and development, especially when caught in a 'debt trap', where they must go on borrowing more and more in order to service old debts (see below). Aid for debt relief helps countries reduce their debt burden and releases resources that can be used for poverty reduction and economic growth and development.

Factors that limit the effectiveness of Official Development Assistance (ODA)

A number of factors limit the effectiveness of aid as a mechanism for achieving economic and human development and poverty alleviation. The most important of these include the following.

Tied aid

One of the most important limitations of the effectiveness of ODA funds is the practice of *tied aid*, whereby donors make the recipients of aid spend all or a portion of borrowed funds to buy goods and services from the donor country. It occurs only in the context of bilateral (not multilateral) aid, and gives rise to several serious disadvantages:

- Recipient countries cannot seek lower price alternatives for the goods and services they are forced to buy from the donor country, so recipients of tied aid face higher than necessary import costs.
- Having to buy specific goods and services from the donor country often results in buying inappropriate, capital-intensive technologies.
- Those who benefit from tied aid are usually large firms in developed countries whose goods and services the recipient countries are forced to buy. This is a kind of support for industry of developed countries, occurring at the expense of poor country development objectives.

See Real world focus 20.2.

Conditional aid (conditionality)

Most donors of ODA impose numerous conditions that must be met by the recipients of aid. Donors see these conditions as a mechanism for forcing developing countries to make important policy changes, as well as for ensuring that aid funds are used effectively. The kinds of conditions vary from requiring the recipient to pursue policies to achieve a greater market orientation (such as privatisation, elimination of trade barriers, etc.), to forcing the recipient to accept particular projects that the donors decide on. Conditional lending creates disadvantages for developing countries. Donors do not pay sufficient attention to the preferences of the government or of the population groups the project is intended to benefit. Policy prescriptions by donors may be incorrect; they may not fit in with the government's development strategy and priorities; and they may weaken the recipient government's authority and accountability to its citizens.

Aid volatility and unpredictability

The flow of aid funds (particularly bilateral flows) into developing countries is volatile (unstable) and unpredictable. This is partly due to changing volumes of aid in donor budgets, and changing donor priorities on how to allocate aid funds. This makes it difficult for recipient governments to implement policies that depend on aid funds, as they cannot be sure if and when funds will be available to undertake necessary investments and activities. In very poor countries that depend heavily on aid for provision of basic services (such as education or health care services), disruptions in aid flows can have very serious effects on the welfare of the population groups affected by the aid cuts.

Uncoordinated donors

In any recipient country there are usually large numbers of donors (bilateral and multilateral) who finance uncoordinated activities, giving rise to numerous inefficiencies in the use of aid resources. Sometimes the numbers of aid-funded projects are in the hundreds. Lack of co-ordination of such projects results in overlapping and duplication of some projects, inconsistencies between other projects, and the lack of coherence in the entire aid effort.

Aid may substitute for rather than supplement domestic resources

Aid resources are intended to supplement insufficient domestic resources. A possible danger is that governments in recipient countries may use aid funds to substitute for domestic resources, and not make enough effort to increase domestic revenues through taxation. The evidence on this issue is mixed; whereas some countries have been unable to increase tax revenues in spite of growth, others have succeeded in increasing tax revenues even as aid increases rapidly.

Aid may not reach those most in need

Aid resources are not allocated on the basis of the greatest need for poverty alleviation. Donors do not allocate aid resources according to country needs, focusing instead on promoting their own interests. As noted by the US Congressional Research Service, aid 'can act as both carrot and stick and is a means of influencing events, solving specific problems and projecting US values'.[32] In addition, recipient country governments may not be genuinely committed to poverty alleviation; they may lack the necessary expertise to design and implement poverty alleviation policies; tied aid may favour projects that are not appropriate for poverty alleviation; donors

[32] https://www.theguardian.com/commentisfree/2018/sep/02/as-a-system-foreign-aid-is-a-fraud-and-does-nothing-for-inequality

may select projects that are not the most effective from the point of view of poverty alleviation.

Aid may be associated with corruption

Corruption involves misuse of aid funds by recipient countries, and is a key problem associated with the provision of aid. Corruption is a reflection of the degree of transparency and accountability in public affairs, and tends to be more prominent the lower the *per capita* income of a country.

The quantity of aid and poverty alleviation

SDG 17.1 (goal 17, target 2) states 'Developed countries to implement fully their official development assistance commitments, including the commitment by many developed countries to achieve the target of 0.7 per cent of ODA/GNI to developing countries . . .'

Donors have repeatedly promised to allocate 0.7% of their GNI for ODA, however only a few meet this target. Since those that do not are among the larger and wealthier donors, it means that overall ODA funds are far less than the target amount. If rich countries fail to follow through on their commitments, developing countries will be unable to make the investments in health, education and infrastructure needed to improve welfare and support the economy on the scale required to achieve the SDGs.

REAL WORLD FOCUS 20.2

Tied aid

The OECD defines *tied aid* as loans or grants offered 'on the condition that it be used to procure goods or services from the provider of the aid'. It has been unsuccessfully advocating for the untying of aid since 2001. It claims that the cost of a development project can increase by 15%–30% due to the tying of aid, preventing 'recipient countries from receiving good value for money for services, goods, or works'.

The OECD notes that untied aid increased from 41% in 1999–2001 to 79% in 2018. However these figures are disputed by the European Network on Debt and Development, which claims that if the informal tying of aid is included almost half of total aid is tied. The worst offenders are the United States with 95% tied aid, Australia with 93% and the United Kingdom with 90%. According to a senior officer,

> *'The procurement of goods and services accounts for almost half of the aid spent by governments around the world according to our estimates. So it is a huge proportion of aid budgets, totalling $55 billion in 2015. Many of these donors regularly acknowledge that untied aid is the way to achieve maximum impact in the fight against poverty and inequalities. However, our report shows that they often fail to practice what they preach.'*[33]

In response to OECD recommendations that member countries untie aid, Japan responded by arguing that 'tied aid is more likely to receive public support in donor countries . . . which in turn helps us increase the support of public funds towards development.'[34]

Figure 20.3: Port Sudan. A docker unloads a bag of sorghum (cereal) from a ship carrying humanitarian aid supplies provided by the US aid agency USAID; the shipment will be distributed to over a million Sudanese in need of assistance

Applying your skills

Research a developing country of your choice that is highly dependent on foreign aid. Investigate whether any of this is tied aid, explore the consequences.

Sources: *https://opinion.inquirer.net; https://eurodad.org; www.devex.com*

[33] https://eurodad.org/Entries/view/1546952

[34] https://www.devex.com/news/oecd-pushes-fix-on-untied-aid-japan-pushes-back-93552

Advantages of NGO: why NGOs are growing in importance

More and more bilateral and multilateral donors of ODA are channelling their funds through NGOs because of their ability to perform some functions better than developing country governments. The reasons for better performance include:

Strong anti-poverty orientation of activities

NGO activities are for the most part concerned with reaching poor people and helping them emerge from their poverty. Governments often have difficulties in reaching the very poor; NGOs have an advantage by working very closely with communities of poor people and responding to their particular needs as these arise in their own particular economic, social and environmental conditions.

Working closely with project beneficiaries

One of the strongest advantages of NGOs is that they work closely with their beneficiaries, involving local people in the design and implementation of development projects. Involvement by local people allows them to participate in deciding what problems should be addressed and how they should be solved, and gives them a sense of ownership and commitment to the project, contributing greatly to success.

Contributing to democratisation, advocacy and raising public awareness and support

Such participatory practices contribute to a process of democratisation, which can be important in countries that do not have democratic institutions. Poor people usually lack political voice and representation, and their concerns are not heard at higher government levels. NGOs play an important leadership role in acting as advocates on public policy issues, and ensuring that poor people's concerns are heard.

Offering expertise and advice

International NGOs accumulate experience from a variety of countries and local settings, many of which may be relevant and transferable to similar settings in other countries. They recruit experts in a variety of areas in accordance with need, and the experts are highly motivated out of a strong commitment to the objectives of the NGO with which they are affiliated.

Ability to be innovative in pursuit of solutions

Unlike governments, which often take a uniform approach to problems, NGOs, by working closely with their beneficiaries, can be more creative and innovative in devising solutions to very specific problems that arise in local settings.

NGOs have a greater freedom than governments to use their expertise and technical knowledge to assess problems independently and arrive at suggestions for solutions. NGOs also enjoy more freedom because their activities are not subject to the conditions often imposed by donors of aid (conditionality); and they are not subject to the restrictions associated with tied aid.

Enjoying the trust of beneficiaries

Poor people are often highly suspicious and mistrusting of government officials and administrators, feeling at best neglected and at worst exploited. NGOs sometimes enjoy greater trust than governments, because of their close relationship with project beneficiaries, and their commitment to solving problems at grassroots level.

Criticisms of NGOs

Small size and weakness of many NGOs

NGOs may be too small and weak to be able to play an important role as agents of change and development. They often have limited resources, and may face difficulties in attracting skilled personnel, so that the effectiveness of their projects may be limited.

Possible loss of independence due to growing dependence on governments and aid agencies for funding

One of the potential strengths of NGOs is their ability to act independently, free of constraints imposed by governments, aid agencies, and bilateral and multilateral donors. However, as they become more and more dependent on these outside sources for their funding, they may lose their independence if they are forced to conform to the demands of funders.

NGOs may attract the best qualified personnel away from government

The growing role of NGOs in development creates a demand for technical experts and personnel that may deprive governments of scarce highly qualified personnel, as NGOs are often in a position to offer higher salaries and benefits than the government.

Challenge to state authority

Whereas governments generally welcome NGOs that complement their activities in poverty alleviation, they often dislike the advocacy role taken on by many NGOs, which may conflict with government policy or question its authority.

The consensus view on NGOs overall is favourable. However, NGOs must act in partnership with governments, and must not be considered to be a replacement of government. Governments have crucial roles to play in the development process, which NGOs, even under the most favourable circumstances, cannot possibly undertake. Governments are essential for establishing an overall policy framework for the economy, including a framework for sustainable development; for providing a legal, institutional and regulatory framework for the economy; for pursuing policies to ensure economic stability; and for correcting market failures.

An example of an NGO, the Grameen Bank, which provides credit to poor people in Bangladesh, is discussed in Real world focus 20.3.

Debt relief

SDG 17.4 (goal 17, target 4) states 'Assist developing countries in attaining long-term debt sustainability through coordinated policies aimed at fostering debt financing, debt relief and debt restructuring'.

In Chapter 19 we saw why indebtedness is a major barrier to growth and development. In 1982, following the buildup of large amounts of debt, the international community stepped in with a series of measures to prevent developing country defaults. These measures included debt restructuring, involving new loans by commercial banks on better terms, such as granting new loans that were stretched out over longer periods at lower interest rates. The loans were used to pay off some of the old loans, and therefore ease the pain of having to service the debts. The IMF gave loans that would help cover large and growing current account deficits. The loans were *conditional* in that they were made only if the borrowing country government agreed to pursue policies prescribed by the IMF including tight fiscal and monetary policies, liberalisation policies and market-based supply-side policies (see above). The World Bank also made conditional loans, which also forced the borrowing country government to pursue economic and trade liberalisation policies to qualify for receiving a loan.

In 1996, the World Bank and IMF began the Heavily Indebted Poor Countries (HIPC) Initiative, intended to provide debt relief to some highly indebted poor countries by cancelling a portion of their external (foreign) debts. The objective was to ensure 'that no poor country faces a debt burden it cannot manage'.[35] In 2005, this was supplemented by the Multilateral Debt Relief Initiative (MDRI) which provides 100% debt relief for debts by the World Bank, IMF and other multilateral institutions).

To qualify for debt cancellation, countries must:

- have a *per capita* GNI below a particular level
- have a debt level that cannot be sustained (i.e. they must be in a debt trap)
- show evidence that they are following certain elements of IMF and World Bank policies (such as cutting government expenditures and liberalising their markets)
- commit themselves to pursuing a poverty reduction strategy.

As of 2019, there were 39 countries eligible to receive HIPC assistance, of which 36 were full debt relief.

Debt relief under the HIPC frees up resources that can be spent for development purposes, since the country's savings from debt reduction must be spent on projects that attack poverty. These include development of rural infrastructure, providing health services and education, creating new jobs, and providing family planning services. In addition, debt reduction is made part of a broader development effort which ensures that more money is directed to the needs of the poor than just the savings from debt reduction. The result is that spending on health, education and other social services is on average about five times the debt-service payments that are saved.

The HICP Initiative is considered to be a welcome step in the direction of solving the debt problem, but has been criticised for several reasons:

- Some of the bilateral creditors have not provided any relief and the rate of delivery of funds remains low. Their participation is voluntary so they cannot be forced to provide funds.
- The programme takes effect too slowly, risking that the benefits of debt relief may follow too slowly to be of much use to the countries.
- Some measures that are imposed as conditions for a country to qualify are too severe, including for example, charging fees for schools and hospitals, privatising key public enterprises such as

[35] https://www.imf.org/en/About/Factsheets/Sheets/2016/08/01/16/11/Debt-Relief-Under-the-Heavily-Indebted-Poor-Countries-Initiative

electricity and telephone, reductions in government expenditures that reduce the provision of social services and infrastructure.

- There are many other countries that are highly indebted but which have not been included in the HIPC Initiative; these countries, whose debt situation is considered to be more manageable, are not poor enough or indebted enough to qualify for assistance, yet still suffer the consequences of high levels of debt, but unable to benefit from debt relief.

TEST YOUR UNDERSTANDING 20.6

1 Define foreign aid, using the concept of concessional flows.

2 Distinguish between
   a ODA and NGOs, and
   b humanitarian aid and development aid, explaining in each case what these consist of.

3 Explain some reasons why foreign aid may be important to the growth and development of poor countries.

4 a Explain the meaning of tied aid and the reasons it limits the effectiveness of aid.
   b Describe some other factors that limit aid's effectiveness.

5 a Outline the kinds of activities NGOs support in developing countries.
   b Suggest reasons why more and more ODA is channelled through NGOs.
   c Discuss strengths and weaknesses of NGOs.

6 Discuss arguments
   a in favour of foreign aid, and
   b against foreign aid.

7 Explain
   a the meaning of debt relief, and
   b some advantages of debt cancellation for heavily indebted countries. (You should refer to Chapter 19.)

8 Discuss some criticisms of the HICP Initiative.

20.7 Multilateral development assistance

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and evaluate the roles of (AO3)
  - the World Bank
  - International Monetary Fund (IMF)

Multilateral development assistance involves lending to developing countries on *non-concessional* terms, in other words with rates of interest and repayment periods determined in the market.

There are a number of major multilateral lenders to developing countries such as:

- multilateral development banks, which lend in order to support economic growth and development, including the:
  - World Bank
  - African Development Bank
  - Asian Development Bank
  - Inter-American Development Bank
  - European Bank for Reconstruction and Development
- the International Monetary Fund (IMF), which lends in order to alleviate external payments difficulties.

Lending by both multilateral development banks and by the International Monetary Fund differ from commercial bank lending because they are involved with lending for economic development or international financial stability, rather than for commercial or profit reasons.

The World Bank

The **World Bank** is a development assistance organisation that extends long-term loans to developing country governments for the purpose of promoting economic development and structural change. It was established in 1944, at the end of the Second World War, as part of an effort to help reconstruct Europe. Its activities were extended to developing countries from the late 1950s when European reconstruction was completed. It is composed of 189 member states that are its joint owners. It consists of two organisations:

- the International Bank for Reconstruction and Development (IBRD), which lends on non-concessional (i.e. commercial) terms to *middle income* developing countries, therefore its activities and lending do not form part of foreign aid; by far the greatest part of World Bank lending for development purposes (about 75%) is offered by the IBRD, and for this reason the World Bank, for the most part, is not considered to be an aid agency.
- the International Development Association (IDA), established in 1960, which has similar activities to the IBRD but extends loans to *low income countries* on highly concessional terms.

The IBRD and IDA are complemented by three additional organisations, that focus mainly on private investments in developing countries.[36]

The importance of the World Bank as a development assistance organisation lies mainly in its role as a lender of funds to governments, and therefore focuses on the roles of the IBRD and IDA.

Brief history of World Bank activities

In the early years of its involvement with developing countries, the World Bank focused on lending for the development of infrastructure, such as energy, transport, telecommunications and irrigation.

By the early 1970s, the World Bank had turned its attention towards poverty alleviation. It grew enormously through an expansion of its funding and technical personnel, and greatly stepped up its lending to developing countries. It redirected a portion of its lending towards poverty alleviation, promising to help the poorest 40% of developing country populations through projects focusing on water supplies, sanitation, education, health, employment, and more.

At the end of the 1970s and in the early 1980s, the Bank's focus changed once more to a new type of lending: *structural adjustment loans (SALs)* intended to change the course of policy-making in developing countries by reducing government intervention and promoting competition and the role of markets. It was believed that a strong market orientation of the economy would help developing countries expand their exports and increase their rates of growth.

Loans were intended to provide assistance in areas like the removal of price controls; interest rate liberalisation (freeing up of interest rates); trade liberalisation (lowering and eliminating tariff and other barriers to trade); eliminating restrictions to new foreign direct investments (by multinational corporations); privatisation (aimed at reducing the size of the public sector); deregulation (aimed at increasing the scope of market forces); cuts in government spending (to reduce budget deficits); and others. Acceptance of the measures included in the loans was a condition that had to be met in order for a country to qualify for a loan.

By the 1990s, SALs had come under very strong and widespread criticism because of their negative consequences on developing country economies. (See the section above *The effects of economic and trade liberalisation.*)

Current World Bank activities

Since the mid-1990s, the World Bank has again shifted towards a poverty orientation, and has committed itself to helping countries achieve first the Millennium Development Goals and more recently the Sustainable Development Goals. Its poverty-oriented projects are meant to be environmentally sustainable; they must not give rise to environmental destruction, and whenever possible they must also improve upon the quality of the environment.

In addition, the World Bank has changed its views on the appropriate role of government in economic growth and development. According to the new perspective, poverty alleviation requires intervention by governments in many areas: education, health care, public health, infrastructure (water, sanitation, transport, irrigation, and many others); access to credit by the poor; land

[36] These are the International Finance Corporation (IFC), International Centre for Settlement of Investment Disputes (ICSID), and Multilateral Investment Guarantee Agency (MIGA).

reforms for a more equitable distribution of agricultural land; policies to reverse environmental degradation; policies to help the poor escape the poverty cycle; policies to promote gender equity. (These issues had been ignored by SALs, with negative consequences for income distribution, poverty and the environment.)

The World Bank is also paying increasing attention to the need for institutional development, based on the idea that markets need institutions that provide education and health services; ensure availability of and access to necessary infrastructure (water, sanitation, transport, etc.); provide an effective and equitable taxation system; ensure access to credit by all who need it; secure property and land rights; minimise the possibilities for the exercise of corruption; empower women and other disadvantaged groups; promote appropriate technology development and innovation; give a political voice to the economically weak; ensure and promote competition; and more.

Evaluating the role of the World Bank

Some of the more important issues include the following:

- **Social and environmental concerns.** In the early years it was criticised for implementing socially unsound projects (such as building hydroelectric dams that displace indigenous people), as well as environmentally unsustainable projects (such as building infrastructure that destroys the natural environment and local ecosystem). In recent years, the World Bank has become far more aware of the social and environmental implications of the projects it funds, and currently makes greater efforts to ensure that project objectives are consistent with the SDGs.
- **World Bank governance dominated by rich countries.** The World Bank is owned by its 189 member states; however, voting power in its governance is determined by the size of financial contributions made by each country to the organisation, which are in proportion to the size of each economy, giving far greater power to rich countries. Critics argue that decisions are made without due regard for the needs and wishes of developing countries.
- **Excessive interference in countries' domestic affairs.** Critics argue that the World Bank interferes excessively in the domestic policy affairs of developing countries.
- **Conditional assistance (lending).** Conditional assistance (or conditional lending) refers to the imposition of conditions that must be met by borrowing countries to qualify for a loan. (It is also one of the problems of foreign aid; see above.) The imposition of conditions is a mechanism for inducing policy changes. It is problematic because it deprives countries of control over their domestic economic activities.
- **Inadequate attention to poverty alleviation.** Although the World Bank has in recent years turned its attention to poverty issues, critics argue that it is not doing enough to meet the challenges of extreme poverty in developing countries by not allocating enough funds for loans intended to meet the needed investments in education, health services, and infrastructure (clean water supplies, sanitation, etc.). In addition, it has been criticised for not doing enough in the area of debt relief through the Heavily Indebted Poor Countries (HIPC) Initiative (see above).
- **Excessive focus on market-based supply-side policies**. The World Bank's *World Development Report, 2019* has been criticised by the International Labour Organization for focusing excessively on increasing flexibility in labour markets, which in the ILO's opinion is not necessary, while ignoring the negative effects on workers[37] (see Chapter 13). Further it has been criticised for encouraging land grabs through privatisations and land takeovers that displace poor farmers (see Chapter 19).

The International Monetary Fund

The **International Monetary Fund (IMF)** is a multilateral financial institution that was established jointly with the World Bank in 1944 with the original purpose of lending to countries experiencing balance of payments deficits under the system of fixed exchange rates that existed at the time. Its objectives have changed over the years in accordance with the evolution of the international financial system. At the present time, the IMF is composed of 189 member countries. Its purpose is to oversee the global financial system, follow the macroeconomic policies of its member countries, stabilise exchange rates and help countries

[37] https://www.ilo.org/global/about-the-ilo/newsroom/statements-and-speeches/WCMS_646884/lang--en/index.htm

that experience difficulties making their international payments by extending them short-term loans on commercial (i.e. non-concessional) terms.

Activities of the International Monetary Fund

In the first two decades of IMF's existence, more than half of its lending was to developed countries. Its role in developing countries grew with the debt crisis beginning in the 1970s and 1980s. This was the time when many poor oil-importing countries developed serious balance of payments difficulties as a result of dramatic increases in oil import expenditures. During the 1990s, the IMF expanded its lending to transition economies in central and eastern Europe and the former Soviet Union. Since 2008 its lending has increased significantly to countries around the world as a result of international payments difficulties brought on by the global financial crisis, including some developed countries (Greece, Iceland, Ireland, Portugal).

The loans provided by the IMF usually come with a package of policies that the country must adopt as a condition for receiving the loan (another example of conditionality). These policies, known as *stabilisation policies*, vary from country to country, but typically include the following:

- contractionary monetary policy, through increases in interest rates, intended to lower aggregate demand, reduce the level of economic activity and reduce demand for imports while encouraging inflows of financial capital, thereby helping the balance of payments position
- contractionary fiscal policy, also intended to lower aggregate demand and reduce the level of economic activity, through cuts in government spending (including cuts in provision of merit goods, such as health services, education, infrastructure, etc.) and cuts in food and other subsidies, as well as increases in taxation, and the imposition of fees for schooling and health care services[38]
- currency devaluation or depreciation, intended to discourage imports and encourage exports and help the balance of payments position
- cuts in real wages (i.e. wages after taking into account the impact of price changes) to lower costs of production as a market-based supply side policy
- liberalisation policies, such as eliminating or reducing controls on prices, interest rates, imports and foreign exchange, to promote a free market and free trade environment.

Evaluating the role of the International Monetary Fund

The IMF is far more controversial than the World Bank, and is subject to more intense criticism due to the negative impact on countries resulting from its stabilisation policies. While it shares some of the criticisms against the World Bank, the main criticism focuses on the harshness of the measures imposed on borrowing countries.

- **IMF governance dominated by rich countries.** As with the World Bank voting power in its governance is in proportion to the size of each economy, giving rich countries far greater power in decision-making.
- **Excessive interference in countries' domestic affairs.** Even more than in the case of the World Bank, critics argue that IMF interference in domestic economies is too great.
- **Conditional lending (conditionality).** Countries have been forced to accept harsh conditions running counter to their growth and development objectives.
- **Damaging effects on developing countries.** Stabilisation policies have the impact of lowering economic growth, often creating a recession with increasing unemployment and increasing levels of poverty. Cuts in real wages where wages are low to begin with, cuts in government spending on merit goods and food subsidies on which many poor people depend for their physical survival, the imposition of fees for schooling and health care services among people who cannot afford them, along with the increases in poverty that arise from liberalisation policies, are wholly inconsistent with economic growth and development objectives, with huge human costs.
- **IMF stabilisation policies based on a flawed concept.** Some economists argue that in addition to the human cost, there may be something fundamentally wrong with the IMF's approach. Experience

[38] Countries with serious balance of payments problems also typically face large government budget deficits (an excess of government spending over government revenues), and so an additional objective of these policies is to reduce government spending and increase revenues in order to reduce the size of the budget deficit.

shows that many countries that have tried the IMF programme suffer not only increasing poverty but also low or negative rates of growth, and therefore are unable to 'grow' out of their balance of payments difficulties or external debt problems. (See Real world focus 17.1.)

It should be noted however that in recent years there are indications that the IMF has moderated the harshness of its policies. An article entitled 'Neoliberalism: Oversold?'[39] states

> *'Instead of delivering growth, some neoliberal policies have increased inequality, in turn jeopardizing durable expansion'*

Referring to three 'disquieting conclusions' the article states that

'• The benefits in terms of increased growth seem fairly difficult to establish when looking at a broad group of countries.
• The costs in terms of increased inequality are prominent . . . .
• Increased inequality in turn hurts the level and sustainability of growth. Even if growth is the sole or main purpose of the neoliberal agenda, advocates of that agenda still need to pay attention to the distributional effects.'

The last point is confirmed by a series of studies by the IMF noted in Chapter 12 (Section 12.4). It therefore appears that the IMF may be rethinking its policies.

TEST YOUR UNDERSTANDING 20.7

1 The lending activities of multilateral organisations such as the World Bank and IMF are sometimes referred to as 'foreign aid'. Outline why this is mostly an incorrect use of the term 'foreign aid'.

2 a Explain the role of the World Bank in developing countries.

b Outline why many of its lending programmes have come under criticism.

c Evaluate the role of the World Bank as a development assistance organisation.

CONTINUED

3 a Describe the role of the International Monetary Fund.

b Explain why it is unpopular in countries to which it lends.

c Evaluate the role of the International Monetary Fund in its aim to assist countries with balance of payments difficulties.

20.8 Institutional change

LEARNING OBJECTIVES

After studying this section you will be able to:

- define all the terms appearing in **orange bold** in the text (AO1)
- explain and evaluate strategies to (AO3)
  - improve access to banking including
    - microfinance
    - mobile banking
  - increase women's empowerment
  - reduce corruption
  - establish property rights and land rights

SDG 16 states 'Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels'.

Microfinance

SDG 1.4 makes reference to *microfinance* as one of items that all men and women, particularly the poor and vulnerable, should have equal access to.

Microfinance, commonly known as *microcredit*, refers to credit (loans) in small amounts to people who do not

[39] https://www.imf.org/external/pubs/ft/fandd/2016/06/ostry.htm

ordinarily have access to credit. 'Micro' is the Greek word for 'small' (μικρό), and refers to the small amounts of the loans, the very small size of businesses or activities that are financed by the loans ('micro-enterprises') and the short repayment periods involved.

Microcredit is delivered to poor people through *microfinance institutions (MFIs)*, which include a wide variety of organisations such as credit unions, financial non-governmental organisations (NGOs), informal savings and loan groups, and even some commercial banks with special programmes for the poor.

Microcredit schemes first appeared in the 1970s as isolated, experimental programmes providing tiny loans to groups of poor women lacking collateral, who invested the borrowed funds in micro-enterprises. As there was no collateral for the loans, loan repayment was ensured and enforced by a guarantee provided for each member by every other member in the group. These kinds of programmes were expanded during the 1980s and 1990s, and revealed that the poor (particularly women) were capable of excellent repayment rates, and were able to pay interest rates high enough to allow the MFIs to cover all their costs. Microcredit schemes have evolved and now provide a great variety of programmes depending on the type of MFI and type of borrower.

MFIs lend to a wide variety of borrowers: micro-enterprises (street vendors, carpenters, seamstresses, and many others, most of whom are in the informal economy), as well as landless rural workers, small farmers, female heads of households, pensioners and displaced persons. They tend to target women because women have proved to be more likely than men to repay loans, and also because they are more likely to use the earnings from their investments to improve the family's well-being.

The evidence on microcredit schemes indicates that these have a positive impact on poverty reduction. They result in higher incomes, more stable incomes, as well as improvements in health, nutrition and primary school attendance. They also result in an improved social and economic status of women.

Microcredit schemes reach only a very small proportion of poor people as there are not enough microcredit schemes and MFIs.

Controversial issues in microfinance

Although most development economists believe that microcredit fills an important gap in access to credit by poor people, many also point out that microcredit schemes carry certain dangers that must be addressed:

- **Microcredit schemes may become a substitute for urgently needed government anti-poverty policies.** There may be a danger that microcredit may be viewed as a substitute for other, complementary government policies needed to combat poverty that are the responsibility of the public sector. For example, government has a role to play in the provision of merit goods, including education, health care services and infrastructure. Poor people should not have to borrow to be able to pay for health care and education, and no amount of borrowing by poor people will produce the needed investments in infrastructure (sanitation, clean water supplies, transport, etc.). Similarly, government must provide protection for people who are disabled, physically or mentally ill, displaced (due to war or other conflict) or in other vulnerable groups, as these people are not appropriate clients for micro-credit.
- **Micro credit schemes contribute to the growth of the informal economy.** The micro-enterprises that are created through micro credit operate for the most part in the informal economy, which is unregulated, where workers have no social protection, and where exploitative conditions often prevail. Microcredit encourages the growth of this controversial sector.
- **Some extremely poor and highly unskilled people may be harmed by micro credit.** The poorest of the poor usually lack skills of all kinds, including skills necessary to begin a micro-enterprise such as basic literacy and numeracy skills. In such cases, microcredit may end up burdening some extremely poor people with payments on loans that cannot produce a stable source of income. In view of this risk, some microcredit schemes try to integrate credit with the provision of education so that borrowers will acquire the skills required to use their loans effectively.
- **Interest rates in micro credit schemes are too high.** Interest rates in microcredit schemes tend to be higher than market rates of interest (although they are much lower than in informal credit markets, such as moneylenders). The costs of providing very many, very small loans are higher than the costs of providing fewer large loans. High interest rates are necessary if MFIs are to be able to cover the costs of providing micro credit. Some economists argue that the high interest rates should be subsidised to make repayment easier for very poor people. Yet donor funds are neither sufficient, nor reliable enough, to be able to cover interest rate subsidies and at the same time allow micro credit schemes to keep on growing. A compromise would involve creating programmes that offer subsidised interest rates only for the poorest borrowers.

REAL WORLD FOCUS 20.3

Muhammad Yunus, the Grameen Bank, and the 2006 Nobel Prize for Peace

In 1976, Professor Muhammad Yunus of the University of Chittagong in Bangladesh initiated a project examining whether it would be possible to provide banking services for the rural poor. Known as the Grameen Bank project ('grameen' means 'village' or 'rural' in the Bangla language), it began making loans to very poor people who had no collateral. At that time, many people believed that the poor could not find paying occupations, and would be unable to pay back loans. The highly successful Grameen project contradicted these beliefs and demonstrated the tremendous potential that extension of credit to poor people has as a strategy for alleviating poverty.

Figure 20.4: New York, USA. Muhammad Yunus speaks onstage during the Whitaker Peace and Development Initiative on 27 September 2019

Over the course of a few years, the original Grameen project spread to several districts of Bangladesh, and in 1983 it was made an independent bank devoted to providing loans to rural poor who lacked collateral. From the most modest beginnings, consisting of $27 lent by Muhammad Yunus to a group of very poor people, Grameen Bank now consists of a few thousand branches with millions of borrowers. In 2006, the Nobel Peace Prize was awarded in two equal parts to Muhammad Yunus and Grameen Bank, in recognition of their efforts to promote economic and social development through the extension of microcredit to the very poor. As the Nobel Committee pointed out, Grameen Bank has been a highly innovative programme, and a source of ideas and models borrowed by many institutions around the world involved with microcredit, and has moreover shown that microcredit is a major weapon in the fight against poverty.

As Muhammad Yunus tells us, these are some of the principles on which Grameen Bank's lending is based:

- Credit is viewed as a human right.
- Its mission is to help poor families help themselves to overcome poverty.
- It is targeted towards the poor, particularly women.
- Lending is not based on collateral, or legally enforceable contracts, but rather on trust.
- A main objective is to create self-employment in income-generating activities such as rice husking, machine repairing, purchase of milk cows and goats, simple manufacturing such as pottery, weaving, garment sewing, simple storage, marketing and transport systems, as well as housing.
- High priority is given to the formation of human capital and environmental protection, community organisation and development, as well as the development of leadership and financial management skills.

It is estimated that those who have benefited most from Grameen borrowing are the landless poor, followed by very small landowners. Over time, incomes have risen more for Grameen borrowers than non-borrowers. There is also a shift away from wage labour (where workers are employed by someone else who pays them a wage) and towards self-employment in micro-enterprises. Many women have improved their status, have become less dependent on their husbands, and have become successful micro-entrepreneurs.

Source: *Adapted from the Grameen Foundation website (www.grameenfoundation.org)*

Applying your skills

1. Identify a country where groups of poor people have benefited from microfinance. Identify possible advantages and disadvantages of microcredit.
2. Muhammad Yunus has given numerous interviews over the years about his work in microcredit. Summarise some of his key views based on one or more of these.

Mobile banking

Mobile banking involves the use of mobile telephones to receive or send money and to pay bills. It has been made possible by advances in mobile technology that allowed money services to grow rapidly.

Globally, 53% of adults do not have access to formal financial services. These vary from a high of 80% in sub-Saharan Africa, 65% in Latin America, 68% in the Middle East, 59% in east and southeast Asia, to 58% in south Asia. This should be contrasted with 8% in high income OECD countries.[40]

In most developing countries there are more people who have mobile phones than have bank accounts. The advantages offered by mobile banking include:

- ease of making payments to family and businesses with instant access and no delays in making or receiving payments
- avoidance of having to travel long distances holding cash which may be stolen
- reduced costs of transferring money
- stronger links between relatives or businesses in rural and urban areas
- easier to pay workers delivering aid to geographically remote areas
- easier to get loans, insurance and other services that facilitate opening a business
- easier to stretch out payments for purchases of equipment needed for a business
- helps women expand their range of activities such as leaving subsistence farming and beginning small businesses
- ease of buying inputs for businesses without having to travel long distances to pay for them
- due to all of the above, mobile banking makes a major contribution toward poverty reduction.

Challenges of mobile banking include:

- network problems, causing delays
- cost of the services, which is high in relation to very low incomes especially in rural areas, though generally lower than banking services
- inability of some older people to read, which makes them more susceptible to fraud.

Women's empowerment

SDG 5 states 'Achieve gender equality and empower all women and girls'. While women account for half the world's population, they represent 70% of the world's poor and they earn 10% of the world's income.[41] See Chapters 18 and 19 for further information on the problem of gender inequalities.

Positive externalities (external benefits) of women's empowerment

For many years, economic development and poverty alleviation efforts made no distinction between the sexes with regard to their implications for growth and development. Since the 1980s and 1990s, development economists have realised that women's empowerment has enormous effects on growth and development, extending beyond the women themselves. These external benefits can be thought of as consumption externalities of women's health and education, analysed as standard health and education externalities:

- **Improvements in child health and nutrition and lower child mortality.** Increased education of women has major positive effects on the health of children, because of improved knowledge about health, health services, basic hygiene and nutrition. Increased education of men leads to a smaller improvement in children's health. Also, increases in women's income levels have a greater positive effect on their children's health than increases in men's incomes.[42]
- **Improvements in educational attainment of children.** Mothers have a major influence over the education of their children, and studies show that the more educated the mother, the more educated the children. As in the case of health, increases in men's education have a smaller impact on their children's education. Further, increases in women's incomes also have a greater impact on their children's education than increases in the incomes of men.
- **Quality of human resources.** The impacts of increased education and incomes of women on levels of health and education of their children have enormous cumulative effects on the quality

[40] https://medium.com/@Bank4youGroup/top-9-mobile-money-advantages-da30558cd8e1

[41] https://borgenproject.org/developing-nations-need-womens-empowerment

[42] One explanation is that women, who tend to be more concerned about their children's well-being, have limited, if any, control over how the husband's income is spent; men, on the other hand, are more likely to spend increases in their income on activities outside the home, including purchases reflecting social status.

of human resources in a country that extend over many years, with the potential to affect profoundly the course of economic growth and development, as well as human development. Development policies that focus on improving the education of women also have a major potential to help poor families and communities break out of the poverty cycle.

- **Lower fertility (lower birth rates).** Increased education of women, more and better work outside the home, and higher incomes lead to having fewer children, because of later marriage and greater reproductive choice, and therefore lower population growth, with all its related benefits (see Real world focus 19.1, Chapter 19).

The Nobel Prize-winning Indian economist Amartya Sen views women as 'active agents of change'. Sen writes that the role of women 'is one of the more neglected areas of development studies, and one that is most urgently in need of correction'. Nothing today in the political economy of development is as important 'as an adequate recognition of political, economic and social participation and leadership of women'.[43] Reflecting this idea, the UNDP has developed a composite indicator, the Gender Inequality Index (GII), measuring gender differences in various dimensions (see Chapter18).

In the 1990s, the United Nations Population Fund (UNFPA) identified the following policies to eliminate inequalities between women and men that are still highly relevant today. [44]

- Establishing mechanisms for equal participation and representation by women in the political process.
- Promoting education, skill development and employment while giving top priority to elimination of poverty, illiteracy and poor health.
- Improving women's ability to earn income beyond traditional occupations, while ensuring equal rights in the labour market.
- Eliminating violence against women.
- Eliminating all discrimination against women, including discriminatory practices by employers.
- Making it possible for women to combine childbearing and child rearing with participation in the labour force.

SDG five, 'Achieve gender equality and empower all women and girls', repeats many of the above points, noting that

> *'Providing women and girls with equal access to education, health care, decent work, and representation in political and economic decision-making processes will fuel sustainable economies and benefit societies and humanity at large. Implementing new legal frameworks regarding female equality in the workplace and the eradication of harmful practices targeted at women is crucial to ending the gender-based discrimination prevalent in many countries around the world.'*[45]

Reducing corruption

SDG 16.5 (goal 16, target 5) states 'Substantially reduce corruption and bribery in all their forms'.

The problem of corruption as a barrier to growth and development was explained in Chapter 19. The International Monetary Fund (IMF, discussed above) notes that the fight against corruption 'requires political will to create strong fiscal institutions that promote integrity and accountability throughout the public sector'.[46] The following policies are recommended:

- Develop high levels of transparency and independent external scrutiny which allows audit agencies and the public to provide supervision. For example Columbia, Costa Rica and Paraguay use an online platform where citizens can monitor physical and financial progress of public investment projects.
- Reform institutions of tax administration.
- Build a professional civil service, based on transparent, merit-based hiring and pay.
- Focus on areas where there is a higher risk of corruption, such as procurement, revenue administration and natural resource management.
- Cooperate with other countries to make it more difficult for corruption to take place across borders. For example more than 40 countries made it a crime for companies to pay bribes to secure business abroad.

[43] Amartya Sen (1999) *Development as Freedom*, Oxford University Press, p. 203. We encountered Amartya Sen in Chapter 18, Section 18.2 in connection with the concept of *human development*.

[44] https://www.unfpa.org/resources/issue-7-women-empowerment

[45] https://www.un.org/sustainabledevelopment/gender-equality

[46] https://blogs.imf.org/2019/04/04/tackling-corruption-in-government

The World Bank tries to help countries fight corruption by:[47]

- Establishing institutions and incentives to prevent corruption from occurring.
- Creating mechanisms that discourage corruption by providing penalties and sanctions.
- Influencing the development of perceptions of the type of governance needed for long term efforts to fight corruption.

Property rights and land rights

SDG 1.4.2 (goal 1, target 4, indicator 2) states 'Proportion of total adult population with secure tenure rights to land, with legally recognized documentation and who perceive their rights to land as secure, by sex and by type of tenure'. This is a measure of equal rights to economic resources, particularly by the poor and vulnerable (target 4).

Property rights and land rights were discussed in Chapter 19. The establishment of property rights that takes the form of titling to property is important for growth as it encourages investment and facilitates credit that allows investment to increase. However, whereas titling can be successful in urban areas it may be problematic in developing countries particularly in connection with land. As we have seen the establishment of property rights based on titles will not provide the necessary legal protection for land based on custom or communal use. Yet the ability of communities to secure land rights is crucially important to sustainable development.[48] Secure land rights:

- contribute to food security as they improve sustainable land use, improve access to credit, and increase productivity of small farmers
- lead to lower rates of deforestation
- preserve diverse food cultures and biodiversity
- support indigenous peoples and improve their economic and social status
- contribute to gender equality when granted to women
- contribute to poverty reduction.

In response to the urgency of this issue, as well as to the problem of land grabs (see Chapter 19), in 2012 the Food and Agricultural Organization (FAO) of the United Nations, in collaboration with civil society, the private sector and research institutions, established *Voluntary Guidelines on the Responsible Governance of Tenure of Land, Fisheries and Forests in the Context of National Food Security*.[49] According to the Director General of the FAO,

> *'Giving poor and vulnerable people secure and equitable rights to access land and other natural resources is a key condition in the fight against hunger and poverty. It is a historic breakthrough that countries have agreed on these first-ever global land tenure guidelines. We now have a shared vision. It's a starting point that will help improve the often dire situation of the hungry and poor.'*[50]

In order to secure land use rights, it is necessary to record individual and collective tenure rights of the state and public sector, the private sector, indigenous peoples and other communities. This means there must be a system of recording, maintaining and publicising tenure rights and responsibilities, including who holds the rights to land, fisheries or forests. The Guidelines set out the best practices countries should follow for the registration and transfer of land use rights, including rights of indigenous communities and mechanisms for resolving disputes.

In addition, there should be anti-eviction legislation, along with access to legal advisors so that formally recognised rights of poor farmers are supported. In cases where extreme poverty is linked with landlessness, governments should consider agrarian reforms aiming to redistribute land to ensure more equitable access.[51]

---

[47] https://www.worldbank.org/en/topic/governance/brief/anti-corruption

[48] https://blogs.worldbank.org/voices/7-reasons-land-and-property-rights-be-top-global-agenda https://www.oxfam.org/en/why-indigenous-and-community-land-rights-matter-everyone

[49] http://www.fao.org/3/i2801e/i2801e.pdf

[50] http://www.fao.org/news/story/en/item/142587/icode

[51] https://journals.openedition.org/poldev/2026

TEST YOUR UNDERSTANDING 20.8

1 a Define microfinance schemes, and discuss their objectives.
  b What are some advantages and disadvantages of micro credit?
2 Use an *AD-AS* diagram to show the effects of an increase in credit leading to an increase in consumption and investment spending.
3 Explain how microfinance can help people escape the poverty cycle.
4 Discuss policies that can contribute to economic growth and development in the areas of
  a increased women's empowerment,
  b reducing corruption, and
  c property rights and land rights.

20.9 Strengths and limitations of government intervention versus market-oriented approaches

LEARNING OBJECTIVES

After studying this section you will be able to (AO3):

- discuss the strengths and limitations of market-oriented and interventionist strategies to promote economic growth and development

We have seen in this chapter that during much of the 20th century, many countries around the world saw significant increases in government intervention. From the 1980s, there was a shift in most countries in the direction of less government intervention and a stronger emphasis on markets. By the early 2000s, it had become apparent that neither the extreme of very strong government intervention, nor the extreme of a highly free market orientation, is appropriate for the conditions of developing countries. Attention of policy-makers therefore turned toward finding an appropriate mix of interventionist and market-based policies. This raises the questions: how much should governments intervene in developing countries, and what are the appropriate roles of markets and intervention?

Strengths and weaknesses of market-oriented policies

Market-oriented policies we have studied include:

- market-based supply-side policies, including:
  - policies encouraging competition (deregulation, privatisation and anti-monopoly regulation)
  - labour market reforms
  - incentive-related policies
- trade liberalisation
- freely floating exchange rates.

Strengths

Market-oriented policies are based on the idea that free markets, working under competitive conditions, offer a method to answer the *what to produce* and *how to produce* questions of resource allocation in the best possible way. With market-determined prices working as signals and incentives, markets co-ordinate the countless independent decisions of consumers, firms and resource owners, allowing social surplus to be maximised, thus achieving allocative efficiency.

Therefore, policies encouraging competition, such as deregulation, privatisation and anti-monopoly regulation, which work by freeing market forces and making markets more competitive, are intended to result in greater efficiency in production, lower prices and improved quality, and a better allocation of resources, as well as economic growth and improved economic well-being.

Labour market reforms similarly promote free market forces in labour markets, allowing the allocation of resources to improve. Incentive-related policies, involving adjustments to various types of taxes, are intended to work by improving the incentives to work, innovate and invest, thus making the signalling and incentive functions of the price mechanism more effective, again improving the allocation of resources and also allowing for economic growth.

Trade liberalisation is based on the same ideas, and has the same intended benefits. The elimination of trade barriers and the opening up of countries to free trade has the effect of making markets much larger than they would be with trade barriers. The result of larger free markets is to increase competition, increase efficiency in production, lower prices and improve quality, increase consumer choice, improve the allocation of resources, and allow for greater economic growth.

Freely floating exchange rates are simply another aspect of the price mechanism of free markets. A market-determined exchange rate is one that reflects the forces of supply and demand for a currency, and therefore can effectively carry out the signalling and incentive function of prices (here applied to the 'price' of a currency) for international transactions of all kinds. Just as a market-determined price of a good 'clears' the market, so too a freely floating exchange rate automatically adjusts to excess demand or supply of a currency, bringing about a balance in the balance of payments and offering greater flexibility to policy-makers to pursue policies needed domestically.

Weaknesses

Market failure

Market-oriented strategies cannot deal with market failures. This is of special importance in many developing countries where market failures of all kinds are more widespread. The market failures that are of particular importance are:

- common pool resources and negative environmental externalities (of production and consumption)
- insufficient provision of merit goods, including education, health care and infrastructure, such as sanitation, clean water supplies, road and transport systems, irrigation, power supplies, etc.
- failure to provide public goods.

All of the above cannot be dealt with by market-oriented policies.

Weak institutional framework

Developing countries need a stronger legal system with enforcement of legal contracts and effective legal recourse, a more effective taxation system, a better-developed banking system, a more effective system of property and land rights, as well as good governance. Market-oriented policies cannot improve institutions.

Insufficient credit for poor people

Poor people do not have access to credit, as the market working on its own does not allow poor people with no collateral and seeking very small loans to acquire the credit they need. This results in lower investment possibilities, greater poverty and poorer income distribution, as well as the inability to escape the poverty cycle.

Income inequalities and poverty

Some market-oriented policies lead to increased income inequalities. The loss of protection of workers resulting from labour-market reforms, and increases in unemployment resulting from some policies to increase competition and reduce the role of the government in the economy, as well as trade liberalisation, which often involves the closure of firms, often result in increases in income inequalities and greater poverty. In addition, the inability of certain groups of people to take advantage of opportunities opened by trade and market liberalisation can also lead to increasing income inequalities.

Inability to alleviate poverty

Poverty alleviation requires policies that redistribute income and wealth including changes in taxation, transfer payments, the imposition of minimum wages, and increased provision of merit goods (education health services, infrastructure) which cannot be carried out by market-oriented policies. Such policies also cannot help people or communities break out of the poverty cycle.

Inability to empower women

Market-oriented policies cannot assist with the empowerment of women because this is an issue that requires provision of merit goods as well as legislation that only governments can undertake.

Informal economy

Market-oriented policies may lead to the growth of the informal economy as workers lose their jobs in the formal economy due to privatisations or reduced size of government or labour market policies that increase worker insecurity.

Questionable effects on economic growth and development

Market and trade liberalisation working on their own may not lead to improved export performance and greater economic growth and development in some countries. According to the evidence, the countries that are better able to take advantage of opportunities offered by trade and market liberalisation are those that have already developed an industrial base, and are therefore better able to withstand the competition arising from the elimination or reduction of trade barriers. Low-income countries tend to perform the worst, because they can least withstand the competition with larger foreign firms, and this sometimes leads to a weakening of their industry together with increased unemployment, poverty and growth of the informal economy.

The withdrawal of government from provision of merit goods that often comes with market liberalisation has negative effects on economic and human development.

Strengths and weaknesses of interventionist policies

Strengths

Interventionist policies are based on government intervention in markets intended to correct market deficiencies and create an environment in which markets can work more effectively. The strengths of interventionist policies include their potential to contribute to the following.

Correcting market failures

Governments have a major role to play in the correction of market failures. This includes policies that try to:

- correct negative environmental externalities of production and consumption and overuse of common pool resources
- provide public goods as well as merit goods that are underprovided by the market due to positive consumption externalities; this involves investments in human capital (health and education) and investments in infrastructure.

Investment in human capital

Investment in human capital (education and health) was noted above in connection with market failures. Education and health have significant external benefits, thus calling for government intervention (such as direct provision) that increases the consumption of both. Education and health are major factors behind increases in productivity that contribute to economic growth, and they also directly lead to greater economic and human development. Investment in human capital also forms a part of industrial policies, discussed below.

Provision of infrastructure

The provision of infrastructure also forms part of policies to correct market failures and includes a broad range of goods and services, also with significant positive externalities. As a type of physical capital, it includes water supplies, sanitation and sewerage, power, communication, transportation, roads, irrigation, and many others. All these play a very important role in encouraging economic growth, as well as making possible economic and human development. They increase productivity, and make a direct contribution to improved standards of living. Therefore, there is a strong role for governments in order to ensure the provision of the appropriate kinds of infrastructure, with the appropriate access by the population.

Development of stronger institutions

Governments can take action to build institutions that enable markets to operate more effectively, including legal, taxation, banking and property and land rights systems.

Redistributing income and reducing poverty

Tax policies, transfer payments and minimum wages, which help ensure that everyone in a society receives at least a minimum income essential to satisfy basic needs, can only be undertaken by the government.

Promotion of gender equality

The government can pursue policies that promote women's empowerment and gender equality in education and health, the labour market, inheritance and property rights, and access to credit.

Industrial policies

Industrial policies are interventionist supply-side policies (see Chapter 13) that include support for small- and medium-sized businesses as well as protection of infant industries (such as through tariffs or subsidies) to help developing countries in the early stages of their industrialisation. They include government support of appropriate technology transfer from developed countries and the establishment of a research and development capability.

Industrial policies were a key factor behind the success of the Asian Tigers. They can play an important role in helping developing countries develop their industries and diversify into higher value-added activities.

Provision of a stable macroeconomic environment

A stable macroeconomic environment includes price stability, full employment, a reasonable budget deficit and a reasonable balance of trade. It requires government intervention through the use of appropriate policies to achieve these objectives. A stable macroeconomic environment is important for ensuring that economic decision-makers (consumers, firms and resource owners) can plan their future economic activities, such as consumption, investment, imports, exports, etc., and is important for investment leading to economic growth.

Weaknesses

Government activities are subject to several weaknesses.

Need for budget funds: opportunity costs and budget deficits and debt

Government policies require use of budget funds, which are usually in short supply in developing countries, due to weaknesses in tax collection as well as low incomes of taxpayers. Any type of government spending has opportunity costs in terms of sacrificed alternatives. In addition, government spending runs the risks of increasing budget deficits and debt.

Excessive bureaucracy and inefficiency

A bureaucracy is an administrative structure of an organisation involving rules that determine how the organisation functions and carries out its tasks. Governments often run into the problem of excessive bureaucracy, meaning there are too many rules governing procedures, red-tape, unproductive workers, high administrative costs and inefficiency. This is a key argument often used in favour of reducing the size of the government sector through privatisation of government-owned enterprises or contracting-out of government activities.

Possible protection of inefficient producers

Certain government policies may lead to protection of inefficient producers, such as trade protection, price floors, subsidies to inefficient firms. This leads to inefficiencies and a waste of resources in the private sector.

Excessive intervention leads to allocative inefficiencies

Too much government intervention, such as in the forms of too many industrial policies, too much intervention in the foreign exchange market, too much trade protection, etc. many lead to allocative inefficiency.

Possible influence of elite groups exerting political pressures

There is a risk that government policies may be influenced by elite groups with too much power over politicians, favouring policies in their own interests rather than the interests of society.

Corruption

We have seen in Chapter 19 that corruption occurs everywhere in the world, but tends to be more important in countries where the legal system, mass media and the system of public administration are weak. It is possible that government or certain government officials may be susceptible to corruption, such as by accepting bribes to carry out certain projects.

Poor governance

Poor governance, referring to the process of designing and implementing policies within a government (see Chapter 19) is also a potential weakness of interventionist policies as these may not be carried out effectively.

TEST YOUR UNDERSTANDING 20.9

1 Discuss some of the main strengths and weaknesses of
   a market-oriented policies, and
   b interventionist policies.
2 Evaluate the view that markets and government intervention should complement each other in developing countries.

20.10 Progress toward meeting selected Sustainable Development Goals

LEARNING OBJECTIVES

After studying this section you will be able to (AO3):

- discuss progress made toward meeting selected SDGs in the context of two or more countries

The Sustainable Development Goals were explained in Chapter 18, where it was also noted that each goal comes with several targets as well as indicators which are used to monitor and measure countries' progress toward achieving the goals and targets.

For example, SDG 2 **Zero hunger** has 7 targets and 14 indicators. One indicator is 'prevalence of undernourishment'. In 2017, there were 821 million people who were undernourished, compared to 784 million in 2015, indicating 'a worrisome rise in world hunger for a third consecutive year'.[52]

SDG 3 **Good health and well-being** has 13 targets and 26 indicators. One indicator is the under-five mortality rate. This fell from 77 deaths (per 1000 live births) in 2000 to 42 in 2015 and 39 in 2017, indicating 'major progress in improving the health of millions of people'.[53]

Similar comparisons can be made for individual countries. See question 5 below under Inquiry and Reflection.

TEST YOUR UNDERSTANDING 20.10

Download the most recent SDG report, select an SDG you are interested in and present your findings.

INQUIRY AND REFLECTION

The following questions will help you reflect on your learning and enhance your understanding of key topics in this chapter. They are suggestions for inquiries that you can undertake on your own or in groups in order to revise the learning objectives of this chapter.

1 Identify a country you are interested in that has succeeded in growing through diversification of its production and exports. Examine the policies it pursued and analyse how diversification contributed to its growth and development.

2 Foreign direct investment is a highly controversial issue. Identify and research a country that has received inward FDI and evaluate its experience, noting the possible positive and negative effects that this has had on its economy. Try to consider the various potential impacts that FDI has on host countries.

3 Some very poor developing countries are almost entirely bypassed by FDI. Identify such a country and research the reasons why it has not attracted any FDI.

4 Foreign aid offers important potential benefits to poor countries. Research a country of your interest that is a recipient of aid and investigate the effects that aid has had on its economy. Consider too any possible negative effects of the aid on the country.

5 Select an SDG and two or more developing countries you are interested in. Go to *the World Bank's DataBank* and select the database called *Sustainable Development Goals*. Choose two or more indicators of your interest, research your countries and present your conclusions regarding relative progress made by the countries.

EXAM STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

[52] https://sustainabledevelopment.un.org/sdg2

[53] https://sustainabledevelopment.un.org/sdg3

Glossary

abnormal profit Arises when total revenue of a firm is greater than its total costs; alternatively when average revenue is greater than average cost, or price is greater than average cost.

absolute advantage Refers to the ability of a country to produce a good using fewer resources than another country, in other words, the ability of a certain amount of resources in a country to produce more than the same resources can produce in another country.

absolute poverty The inability of an individual or a family to afford a basic standard of goods and services, where this standard is absolute and unchanging over time; it is defined in relation to a nationally or internationally determined *poverty line*, which determines the minimum income that can sustain a family in terms of its basic needs.

abuse of market power Occurs when firms engage in activities that restrict competition; see *market power.*

actual growth In the context of the production possibilities (*PPC*) model, it is growth that occurs due to reduction of unemployment or improvement in efficiency of resource use, resulting in a movement of a point inside the *PPC* to a point closer to the *PPC* in the northeast direction.

Adam Smith (1723–1790) a Scottish philosopher often referred to as the 'Father of economics, best known for his idea that the self-interested behavior of decision-makers without government intervention results in competitive markets leading to a more efficient use of resources and increased output; see *laissez faire*.

administrative barriers Trade protection measures taking the form of administrative procedures that countries may use to prevent the free flow of imports into a country; may include customs procedures involving inspections and valuation, controls on packaging, and others. Often considered to be a kind of 'hidden' trade protection.

adverse selection A type of asymmetric information where one party has more information than the other party about the quality of the product being sold. If the seller has more information, such as when selling a used car, the buyer will reduce demand. If the buyer has more information, such as regarding one's health condition when buying health insurance, the seller will reduce supply. The result in both cases is underallocation of resources to the good or service.

aggregate demand The total quantity of goods and services that all buyers in an economy (consumers, firms, the government and foreigners) want to buy over a particular time period, at different possible price levels, *ceteris paribus.*

aggregate demand curve A curve used in macroeconomics showing the relationship between the total amount of real output demanded by the four components (consumers, firms, government, foreigners) and the economy's price level over a particular time period, *ceteris paribus*.

aggregate supply The total quantity of goods and services produced in an economy over a particular time period, at different price levels, *ceteris paribus.*

allocative efficiency An allocation of resources that results in producing the combination and quantity of goods and services mostly preferred by consumers. The condition for allocative efficiency is given by $MSB = MSC$ (marginal social benefit = marginal social cost or $P = MC$ (price is equal to marginal cost); alternatively it is when *social surplus* is maximum.

allocative inefficiency The absence of *allocative efficiency*; when $MSB \neq MSC$ (or $P \neq MC$).

anchoring Part of *behavioural economics*, involves the use of irrelevant information to make decisions, which often occurs due to its being the first piece of information that the consumer happens to come across.

anti-dumping An argument that justifies trade protection policies: if a country's trading partner is suspected of practising dumping (selling a good in international markets at a price below the cost of producing it), then the country should have the right to impose trade protection measures (tariffs or quotas) to limit imports of the dumped good; dumping is illegal according to international trade rules.

appreciation (of a currency) An increase in the value of a currency in the context of a floating exchange rate system or managed exchange rate system (compare with *revaluation*, which refers to an increase in currency value in the context of a fixed or pegged exchange rate system).

appropriate technology Technologies that are well-suited to a country's particular economic, geographical, ecological and climate conditions. Often used in connection with labour-abundant developing countries that require labour-intensive (as opposed to capital-intensive) technologies.

asymmetric information A type of market failure where buyers and sellers do not have equal access to information, usually resulting in an underallocation of resources to the production of goods and services, as parties to a transaction with less access to information try to protect themselves against the consequences of the information asymmetry.

automatic stabilisers Factors that automatically, without any action by government authorities, work toward stabilising the economy by reducing the short-term fluctuations of the business cycle. Two important automatic stabilisers are progressive income taxes and unemployment benefits.

availability A term from *behavioural economics*, refers to use of information that is most recently available, which people tend to rely on more heavily, though there is no reason that this information is any more reliable than previously available information.

average costs Costs per unit of output, or the cost of each unit of output on average. They are calculated by dividing total cost by the number of units of output produced.

average revenue Revenue per unit of output sold, calculated by dividing total revenue by the number of units of output produced.

average tax rate Tax paid divided by total income, expressed as a percentage (i.e. tax paid divided by total income multiplied by 100).

balance of payments A record (usually for a year) of all transactions between the residents of a country and the residents of all other countries, showing all payments received from other countries (credits), and all payments made to other countries (debits).

balance of trade in goods Part of the balance of payments, it is the value of exports of goods minus the value of imports of goods over a specific period of time (usually a year).

balance of trade in services Part of the balance of payments, it is the value of exports of services minus the value of imports of services over a specific period of time (usually a year).

barriers to entry Anything that can prevent a firm from entering an industry and beginning production, as a result limiting the degree of competition in the industry.

base rate See *minimum lending rate*.

behavioural economics A relatively new branch of economics strongly influenced mainly by psychology, but also by sociology and neuroscience based on the idea that human behaviour is far more complex than the assumptions of *rational consumer choice*.

biases A terms from *behavioural economics,* it is a term from psychology that refers to systematic errors in thinking or evaluating; examples include *rules of thumb*, *anchoring*, *framing*, *availability*.

bilateral trade agreement Any trade agreement (or agreement to lower international trade barriers) involving two trading partners, usually two countries. It may also involve a trade agreement between one country and another group of countries when this groups acts as a single unit (such as the European Union). May be contrasted with *regional trade agreement* and *multilateral trade agreement*.

bounded rationality Part of *behavioural economics,* it is the idea that consumers are rational only within limits, as consumer rationality is limited by consumers' insufficient information, the costliness of obtaining information, and the limitations of the human mind to process large amounts of information.

bounded self-control Part of *behavioural economics,* it is the idea that people in reality exercise self-control only within limits, lacking the self-control required of them to make rational decisions.

bounded selfishness Part of *behavioural economics,* it is the idea that people are selfish only within limits; the assumption of self-interested behaviour of the rational consumer cannot explain the numerous accounts of selfless behaviour.

budget deficit Referring usually to the government's budget, it is the situation where government tax revenues are less than government expenditures over a specific period of time (usually a year).

budget surplus Referring usually to the government's budget, it is the situation where government tax revenues are greater than government expenditures over a specific period of time (usually a year).

business confidence A measure of the degree of optimism among firms in an economy about the future performance of firms and the economy; it is measured on the basis of surveys of business managers. Is an important determinant of the investment component of aggregate demand.

business cycle Fluctuations in the growth of real output, or real GDP, consisting of alternating periods of expansion (increasing real output) and contraction (decreasing real output).

capital One of the factors of production, which itself has been produced (it does not occur naturally), also known as 'physical capital', including machinery, tools, equipment, buildings, etc. Other types of capital include 'human capital', or the skills, abilities, knowledge and levels of good health acquired by people; 'natural capital', or everything that traditionally has been included in the factor of production 'land'; and 'financial capital', or purchases of financial instruments such as stocks and bonds.

capital account In the balance of payments, refers to the inflows minus outflows of funds for (i) capital transfers (including such things as debt forgiveness and non-life insurance claims), and (ii) the purchase or use of non-produced natural resources (such as mineral rights, forestry rights, fishing rights and airspace); it is a relatively unimportant part of the balance of payments.

capital expenditures With reference to government expenditures, these include public investments, or the production of physical capital, such as building roads, airports, harbours, school buildings, hospitals, etc.

capital flight The large-scale transfer of privately-owned financial capital (funds) to another country resulting from fear and uncertainty of holding domestic assets.

capital gains tax A tax on profits from financial investments such as stocks and bonds or from buying and selling real estate.

capital transfers A part of the capital account of the balance of payments, they include inflows minus outflows for such things as debt forgiveness, non-life insurance claims, and investment. See *capital account*.

carbon tax A tax per unit of carbon emissions of fossil fuels, considered by many countries as a policy to deal with the problem of climate change.

central bank A financial institution responsible for regulating the country's financial system and commercial banks, and carrying out monetary policy.

ceteris paribus A Latin expression meaning 'other things being equal'; all other variables other than those under investigation are assumed to be constant or unchanging.

change One of the key concepts of this course; change is important in economics in the study of both economic theory as well as in real world events. *One of the key concepts of this course*

choice One of the key concepts of this course; economics is a study of choices, or selecting among alternatives, due to the scarcity of resources. *One of the key concepts of this course*

choice architecture A term in *behavioural economics*, it is the design of particular ways or environments in which people make choices, based on the idea that consumers make decisions in a particular context and that choices of decision-makers are influenced by how options are presented to them.

circular economy The idea that goods should be produced in such a way that they can be repaired rather than thrown out; they would be made out of biological materials so that once discarded they can go back to the biosphere and prevent pollution of the planet.

circular flow of income a flow of income in an economy where the value of output produced is equal to the total income generated in producing that output, which is equal to the expenditures made to purchase that output (see *circular flow of income model*).

circular flow of income model A model showing the flow of resources from consumers (households) to firms, and the flow of products from firms to consumers, as well as money flows consisting of consumers' income arising from the sale of their resources and firms' revenues arising from the sale of their products.

classical economics Economic ideas of the nineteenth century; a main feature was the idea that markets working on their own according to the principles of supply and demand could solve all major economic problem, including unemployment and recession, and allocate resources efficiently.

collective self-governance A solution to the use of common pool resources where the users take control of the resources and use them in a sustainable way; runs counter to the idea of the *tragedy of the commons*. This solution presupposes that the users of the resources can communicate with each other, resulting in rules about the use of the resources along with sanctions for violations of the rules.

collusion An agreement among firms to fix prices, or divide the market between them, so as to limit competition and maximise profit; usually involves firms in oligopoly.

collusive oligopoly Refers to the type of oligopoly where firms agree to restrict output or fix the price, in order to limit competition, increase market power (monopoly power) and increase profits.

command and control Refers to government laws and regulations that everyone must follow.

commercial bank A financial institution (which may be private or public) whose main functions are to hold deposits for their customers (consumers and firms), to make loans to their customers, to transfer funds by cheque (check) from one bank to another, and to buy government bonds; they are regulated by the *central bank*.

common market A type of trading bloc in which countries that have formed a customs union proceed further to eliminate any remaining tariffs in trade between them; they continue to have a common external policy (as in a *customs union*), and in addition agree to eliminate all restrictions on movements of any factors of production within them; factors affected are mainly labour and capital. The best-known common market is the European Economic Community (EEC, precursor of the present European Union).

common pool resources Resources that are not owned by anyone, do not have a price, and are available for anyone to use without payment (for example, lakes, rivers, fish in the open seas, open grazing land, the ozone layer and many more); their depletion or degradation leads to environmental unsustainability.

community surplus See *social surplus*.

comparative advantage Arises when a country has a lower relative cost, or opportunity cost, in the production of a good than another country. Forms the basis of the *theory of comparative advantage*.

competition Occurs when there are many buyers and sellers acting independently, so that no one has the ability to influence the price at which the product is sold in the market.

competitive market A market composed of many buyers and sellers acting independently, none of whom has any ability to influence the price of the product (i.e. no *market power*).

competitive market equilibrium The equilibrium that emerges at the point where the demand curve intersects the supply curve in a free competitive market (where there is no government intervention).

competitive supply In the case of two goods, refers to production of one or the other by a firm; in other words the two goods compete with each other for the same resources (for example, if a farmer can produce wheat or corn, producing more of one means producing less of the other).

complements (complementary goods) Two or more goods that tend to be used together. If two goods are complements, an increase in the price of one will lead to a decrease in the demand of the other.

composite indicator A summary measure of more than one indicator, often used to measure economic development; for example the *Human Development Index* (*HDI*), that measures income, education and health indicators.

concentration ratio A measure of how much an industry's production is concentrated among the industry's largest firms; it measures the percentage of output produced by the largest firms in an industry, and is used to provide an indication of the degree of competition or degree of market power in an industry. The higher the ratio, the greater the degree of market power; see also *market concentration*.

consumer confidence A measure of the degree of optimism of consumers about their future income and the future of the economy; it is measured on the basis of surveys of consumers. Is an important determinant of the consumption component of aggregate demand.

consumer nudge See *nudge*.

consumer price index A measure of the cost of living for the typical household; it compares the value of a basket of goods and services in one year with the value of the same basket in a base year. Inflation (and deflation) are measured as a percentage change in the value of the basket from one year to another.

consumer surplus Refers to the difference between the highest prices consumers are willing to pay for a good and the price actually paid. In a diagram, it is shown by the area under the demand curve and above the price paid by consumers up to quantity purchased.

consumption Spending by households (consumers) on goods and services (excludes spending on housing).

contracting out A practice often undertaken by the government when it makes an agreement (or contract) with a private firm to carry out an activity that the government was previously doing itself.

contractionary fiscal policy Refers to fiscal policy usually pursued in an inflation, involving a decrease in government spending or an increase in taxes (or both). May be contrasted with *expansionary fiscal policy*. See also *fiscal policy*.

contractionary monetary policy Refers to monetary policy usually pursued in an inflation, involving an increase in interest rates, intended to lower investment and consumption spending. May be contrasted with *expansionary monetary policy*. See also *monetary policy*.

corporate income tax Tax on the profits of corporations, which are businesses (firms) that have formed a legal body called a 'corporation' that is legally separate from its owners.

corporate indebtedness The degree to which corporations have debts.

corporate social responsibility The practice of some corporations to avoid socially undesirable activities, such as polluting activities, employing children, or employing workers under unhealthy conditions; as well as undertaking socially desirable activities, such as support for human rights and donations to charities.

cost-push inflation A type of inflation caused by a fall in aggregate supply, usually resulting from increases in costs of production (for example, wages or prices of other inputs), shown in the *AD-AS* model as a leftward shifts of the *AS* curve.

costs of production Payments by firms to obtain and use factors of production in their production process.

credit items In the balance of payments, refer to payments received from other countries, entering the balance of payments accounts with a plus sign; they represent an inflow of foreign exchange into a country.

credit rating An assessment of the ability of a borrower to pay back loans, usually carried out by agencies that are qualified to do this; a high credit rating received by a government means that it is expected to be able to pay back its loans in full and on time without difficulties.

crowding out Refers to the possible impacts on real GDP of increased government spending (expansionary fiscal policy) financed by borrowing; if increased government borrowing results in a higher rate of interest, this could reduce private investment spending, thus reversing the impacts of the government's expansionary fiscal policy.

current account balance The sum of credits plus debits in the current account of the *balance of payments*.

current account deficit Occurs when the current account balance has a negative value, meaning that debits are larger than credits (there is an excess of debits).

current account In the balance of payments, this includes the balance of trade (recording exports minus imports of goods) plus the balance on services (recording exports of services minus imports of services), plus inflows minus outflows of income and current transfers. The most important part of the current account in most countries is the balance of trade.

current account surplus Occurs when the current account balance has a positive value, meaning that credits are larger than debits (there is an excess of credits).

current expenditures In the government budget, refers to government spending on day-to-day items that are recurring (i.e. repeat themselves) and items that are used up or 'consumed' as a good or service is provided. Include wages and salaries (for all government employees); spending for supplies and equipment for the day-to-day operation of government activities (for example, school supplies and medical supplies for public schools and public health care services); provision of subsidies; and interest payments on government loans.

current transfers An item in the current account of the balance of payments, refers to inflows and outflows of funds for items including gifts, foreign aid and pensions.

customs union A type of trading bloc, consisting of a group of countries that fulfil the requirements of a free trade area (elimination of trade barriers between members) and in addition adopt a common policy towards all non-member countries; members of a customs union also act as a group in all trade negotiations and agreements with non-members.

cyclical unemployment A type of unemployment that occurs during the downturns of the business cycle, when the economy is in a recessionary gap; the downturn is seen as arising from declining or low aggregate demand, and therefore is also known as 'demand-deficient' unemployment.

debit items In the balance of payments, refer to payments made to other countries, entering the balance of payments accounts with a minus sign; they represent an outflow of foreign exchange from a country.

debt relief Refers to the cancellation or forgiveness of all or a portion of a country's debt.

debt servicing The payments that must be made in order to repay the principal (the amount of a loan) plus interest.

default choice A term in *behavioural economics*; it is a choice that is made by default, which means doing the option that results when one does not do anything.

deferred consumption Occurs when consumers postpone spending, such as if they expect the price level to fall.

deficit In general, this is the deficiency of something compared with something else. (i) In the balance of payments, a 'deficit' in an account occurs when the credits (inflows of money from abroad) are smaller than the debits (outflows of money to other countries); for example, a deficit in the balance of trade means that the value of exports (credits) is smaller than the value of imports (debits). (ii) In the case of the government budget, a 'deficit' occurs when government revenues are smaller than government expenditures.

deflation A continuing (or sustained) decrease in the general price level.

deflationary gap See *recessionary gap*.

demand Indicates the various quantities of a good that consumers (or a consumer) are willing and able to buy at different possible prices during a particular time period, *ceteris paribus* (all other things being equal).

demand curve A curve showing the relationship between the price of a good and the quantity of the good demanded, *ceteris paribus* (all other things equal); see *demand*.

demand management Policies that focus on the demand side of the economy, attempting to influence aggregate demand to achieve the goals of price stability, full employment and economic growth.

demand-deficient unemployment See *cyclical unemployment.*

demand-pull inflation A type of inflation caused by an increase in aggregate demand, shown in the *AD-AS* model as a rightward shift in the *AD* curve.

demand-side policies Policies that attempt to change aggregate demand (shift the aggregate demand curve in the *AD-AS* model) in order to achieve the goals of price stability, full employment and economic growth, and minimise the severity of the business cycle. In the event of an inflationary or recessionary (deflationary) gap, they try to bring aggregate demand to the full employment level of real GDP, or potential GDP. They can also impact on economic growth by contributing to increases in potential GDP. Consists of fiscal and monetary policies. To be contrasted with *supply-side policies*.

demerit goods Goods that are considered to be undesirable for consumers and are overprovided by the market. Reasons for overprovision are usually that the goods have *negative consumption externalities;* in addition there may be consumer ignorance about the harmful effects.

depreciation (of a currency) Refers to a decrease in the value of a currency in the context of a floating exchange rate system or managed exchange rate system (to be compared with devaluation, which is a decrease in currency value in a fixed or pegged exchange rate system).

deregulation Policies involving the elimination or reduction of government regulation of private sector activities, based on the argument that government regulation stifles competition and increases inefficiency.

determinants of aggregate demand Factors that cause shifts of the aggregate demand curve; include factors that influence consumption spending (C), investment spending (I), government spending (G) and net exports ($X - M$).

determinants of demand See *non-price determinants of demand*.

determinants of supply See *non-price determinants of supply*.

devaluation (of a currency) Refers to a decrease in the value of a currency in the context of a fixed or pegged exchange rate system (to be compared with *depreciation*, which is a decrease in currency value in the context of a floating or managed exchange rate system).

development aid Foreign aid intended to help economically less developed countries with their growth and development efforts; to be contrasted with *humanitarian aid*.

direct taxes Taxes paid directly to the government tax authorities by the taxpayer, including personal income taxes, corporate income taxes and wealth taxes.

discount rate See *minimum lending rate* of central bank.

disinflation Refers to a fall in the rate of inflation; it involves a positive rate of inflation and should be contrasted with *deflation*.

distribution of income Concerned with how much of an economy's total income different individuals or different groups in the population receive, and involves answering the '*for whom*' basic economic question.

diversification Generally refers to change involving greater variety, and is used to refer to increasing the variety of goods and services produced and/or exported by a country; it is the opposite of *specialisation*.

economic development Broad-based rises in standards of living and well-being of a population, particularly in developing countries. It involves increasing income levels and reducing poverty, reducing income inequalities and unemployment, and increasing provision of and access to basic goods and services such as food and shelter, sanitation, education and health care services.

economic growth Increases in total real output produced by an economy (real GDP) over time; may also refer to increases in real output (real GDP) *per capita* (or per person).

economic inequality The degree to which people in a population differ in their ability to satisfy their economic needs; economists focus on inequalities that result mainly from differences in income and wealth.

economic integration Refers to economic interdependence between countries, usually achieved by agreement between countries to reduce or eliminate trade and other barriers between them. There are various degrees of integration, depending on the type of agreement and the degree to which barriers between countries are removed; see *trading bloc*, *free trade area*, *customs union*, *common market*, *monetary union*.

economic well-being One of the key concepts of this course; refers to levels of prosperity, economic satisfaction and standards of living among the members of a society. *One of the key concepts of this course*

economically least developed countries Developing countries with very low incomes that have low levels of human capital and a high degree of economic vulnerability; they are challenged in achieving *sustainable development*.

economics The study of choices leading to the best possible use of scarce resources in order to best satisfy unlimited human needs and wants.

economies of scale Decreases in the average costs of production that occur as a firm increases its output by varying all its inputs (i.e. in the long run); explain the downward-sloping portion of the long-run average total cost curve: as a firm increases its size, the costs per unit of output fall.

efficiency One of the key concepts of this course; involves making the best possible use of scarce resources to avoid waste; may refer to producing at the lowest possible cost, or producing what consumers mostly want (see *allocative efficiency*).

elastic Refers to a high responsiveness of a variable (such as quantity demanded) to a change in another variable (such as price or income); see various elasticities.

elasticity In general, this is a measure of the responsiveness or sensitivity of a variable to changes in any of the variable's determinants. See specific elasticities: *price elasticity of demand*, *income elasticity of demand*, *price elasticity of supply*.

empirical evidence Real-world information, observations and data that we acquire through our senses and experience.

Engel curve A curve that shows the relationship between consumer income and demand for a product; indicates whether a good is normal or inferior.

entrepreneurship One of the factors of production, involving a special human skill that includes the ability to innovate by developing new ways of doing things, to take business risks and to seek new opportunities for opening and running a business. Entrepreneurship organises the other three factors of production (land, labour and capital) and takes on the risks of success or failure of a business.

equality The state of being equal with respect to something; income equality means everyone receives the same income. *One of the key concepts of this course*

equilibrium A state of balance such that there is no tendency to change

equilibrium level of output The level of output (real GDP) where the aggregate demand curve intersects the aggregate supply curve (also known as the 'equilibrium level of income'). Note the distinction between *short-run equilibrium level of output* and *long-run equilibrium level of output*.

equilibrium level of real GDP See *equilibrium level of output*.

equilibrium price The price determined in a market when quantity demanded is equal to quantity supplied, and there is no tendency for the price to change; it is the price that prevails when there is *market equilibrium*.

equilibrium quantity The quantity that is bought and sold when a market is in equilibrium, i.e. when quantity demanded is equal to quantity supplied.

equity One of the key concepts of this course; is the condition of being fair or just; should be contrasted with the term 'equality'. Often used in connection with income distribution, in which case it is usually interpreted to mean income equality (though this is only one possible interpretation of equity).

excess demand In the context of demand and supply, occurs when the quantity of a good demanded is greater than the quantity supplied, leading to a shortage of the good; see *shortage*.

excess supply In the context of demand and supply, occurs when the quantity of a good demanded is smaller than the quantity supplied, leading to a surplus; see *surplus*.

exchange rate The rate at which one currency can be exchanged for another, or the number of units of foreign currency that correspond to the domestic currency; can be thought of as the 'price' of a currency, which is expressed in terms of another currency.

excludable A characteristic of goods according to which it is possible to exclude people from using the good by charging a price for it; if someone is unwilling or unable to pay the price they will be excluded from using it. Most goods are excludable. See also *rivalrous*.

expansionary fiscal policy Refers to fiscal policy usually pursued in a recession, involving an increase in government spending or a decrease in taxes (or both). May be contrasted with *contractionary fiscal policy*. See also *fiscal policy*.

expansionary monetary policy Refers to monetary policy usually pursued in a recession, involving a decrease in interest rates, intended to increase investment and consumption spending. May be contrasted with *contractionary monetary policy*. See also *monetary policy*.

expenditure approach A method used to measure the value of aggregate output of an economy, which adds up all spending on final goods and services produced within a country within a given time period ($C + I + G + (X - M)$). As suggested by the circular flow model, it is equivalent to measurement by the *output approach* and the *income approach*.

expenditure reducing policies Policies that involve reducing expenditures in the domestic economy so as to bring about a decrease in imports in order to correct a current account deficit; they include contractionary fiscal and monetary policies.

expenditure switching policies Policies that involve switching consumption away from imported goods and towards domestically produced goods, in order to correct a current account deficit; include trade protection policies and depreciation.

export promotion Refers to a growth and trade strategy where a country attempts to achieve economic growth by expanding its exports. As a trade strategy, it looks towards foreign markets and is based on stronger links between the domestic and global economies. To be contrasted with *import substitution*.

export subsidy A payment by the government to a producer or exporter per unit of the subsidised good, where the subsidy is paid for each unit of the good that is exported.

exports Goods or services that are sold to other countries.

external balance A situation where the revenues from exports of goods and services are roughly equal to the expenditures on imports of goods and services over an extended period of time

externality Occurs when the actions of consumers or producers give rise to positive or negative side-effects on other people who are not part of these actions, and whose interests are not taken into consideration. Positive externalities give rise to positive side-effects; negative externalities to negative side-effects.

factors of production All resources, or inputs (land, labour, capital, entrepreneurship) used to produce goods and services.

financial account In the balance of payments, refers to inflows minus outflows of funds due to foreign direct investment, portfolio investment, changes in reserve assets and changes in official borrowing.

firm A business involved in production of goods or services; a key decision-maker in a market economy.

fiscal policy Manipulations by the government of its own expenditures and taxes in order to influence the level of aggregate demand; it is a type of *demand-side policy* or *demand management*.

fixed exchange rate Refers to an exchange rate that is fixed by the central bank of a country, and is not permitted to change in response to changes in currency supply and demand; requires constant intervention by the central bank or government.

fixed exchange rate system An exchange rate system where exchange rates are fixed by the central bank of each country. See *fixed exchange rate*.

floating exchange rate An exchange rate determined entirely by market forces, or the forces of supply and demand. There is no government intervention in the foreign exchange market to influence the value of the exchange rate.

floating exchange rate system An exchange rate system where exchange rates are determined entirely by market forces; see *floating exchange rate*.

for whom to produce One of the three basic economic questions; refers to the choice that must be made in response to the question of income or wealth distribution among a population.

foreign aid The transfer of funds or goods and services to developing countries with the main objective to bring about improvements in their economic, social or political conditions. Such transfers must be 'concessional' (loans must be on more favourable terms than a bank would give) and 'non-commercial' (must not be involved with buying and selling activities concerned with making a profit).

foreign direct investment (FDI) Refers to investment by firms based in one country (the home country) in productive activities in another country (the host country) with control of at least 10 per cent of the firm in the host country. Firms that undertake FDI are called *multinational corporations*.

foreign exchange Refers to foreign national currencies, i.e. for any country, it refers to currencies other than its own.

framing A term from *behavioural economics*, it deals with how choices are presented to decision-makers; for example, consumers might be willing to pay a higher price for a faded pair of jeans in a boutique than for the identical pair of jeans in a discount store.

free entry The condition in which firms face no barriers to entering an industry, characteristic of the market structures of *perfect competition* and *monopolistic competition*.

free good Any good that is not scarce, therefore has a zero opportunity cost.

free market economy An 'ideal type' of economy based on the market approach to making economic decisions; involve private sector ownership and decision-making, and price rationing; to be contrasted with a *planned economy*.

free rider problem Occurs when people can enjoy the use of a good without paying for it, and arises from non-excludability: people cannot be excluded from using the good, because it is not possible to charge a price. Is often associated with public goods, which are a type of market failure: due to the free rider problem, private firms fail to produce these goods.

free trade The absence of government intervention of any kind in international trade, so that trade takes place without any restrictions (or barriers) between individuals or firms in different countries.

free trade area (agreement) A type of trading bloc, consisting of a group of countries that agree to eliminate trade barriers between themselves; it is the most common type of integration area, and involves a lower degree of economic integration than a *customs union* or *common market*. Each member country retains the right to pursue its own trade policy towards non-member countries. An example is NAFTA (North American Free Trade Agreement).

frictional unemployment A type of unemployment that occurs when workers are between jobs; workers may leave their job because they have been fired, or because their employer went out of business, or because they are in search of a better job, or they may be waiting to begin a new job; tends to be short term.

full employment (i) In the production possibilities model, refers to maximum use of all resources in the economy to produce the maximum quantity of goods and services that the economy is capable of producing (production possibilities), implying zero unemployment. (ii) In the *AD–AS* model, refers to the *natural rate of unemployment*, or unemployment of labour that prevails when the economy is producing potential output, or real GDP, determined by the position of the *LRAS* curve (when the economy is in long equilibrium).

full employment level of output (real GDP) The level of output (or real GDP) at which unemployment is equal to the natural rate of unemployment; the level of output (real GDP) where there is no deflationary or recessionary gap. Also known as *potential output (potential GDP)*.

gains from trade Refer to the gains from international trade that arise when countries specialise and trade according to their absolute or comparative advantage; may be in terms of increased output or increased consumer or producer surplus.

game theory A mathematical technique analysing the behaviour of decision-makers who are dependent on each other, and who use strategic behaviour as they try to anticipate the behaviour of their rivals. Has become an important tool in microeconomics, often used to analyse the behaviour of oligopolistic firms; is based heavily on the work of American mathematician and economist John Nash.

GDP *per capita* Gross domestic product divided by the number of people in the population; is an indicator of the amount of domestic output per person in the population.

GDP See *gross domestic product.*

Gender Inequality Index (GII) A composite indicator that measures inequalities between the genders in three dimensions: reproductive health, empowerment and labour market participation.

Gini coefficient (Gini index) A summary measure of the information contained in the *Lorenz curve* of an economy, defined as the area between the diagonal and the Lorenz curve, divided by the entire area under the diagonal. The Gini coefficient has a value between 0 and 1; the larger the Gini coefficient, and the closer it is to 1, the greater is the income inequality.

globalisation Refers to economic integration on a global scale, involving increasing interconnectedness throughout the world in many areas (trade, finance, investment, people, technology, ideas, knowledge, communications and culture).

GNI See *gross national income*.

GNI *per capita* Gross national income divided by the number of people in the population; is an indicator of the amount of income in an economy per person in the population.

governance Refers to the way of governing, and the exercise of power in the management of an economy's economic and social resources, in order to achieve particular objectives such as economic growth and development.

government debt Refers to the government's accumulation of budget deficits minus budget surpluses; is the total amount owed by the government to all creditors (lenders); also known as *national debt*.

government intervention The practice of government to intervene (interfere) in markets, preventing the free functioning of the market, usually for the purpose of achieving particular economic or social objectives.

gross domestic product (GDP) A measure of the value of aggregate output of an economy, it is the market value of all final goods and services produced within a country during a given time period (usually a year); may be contrasted with *gross national income (GNI)*.

gross national income (GNI) A measure of the total income received by the residents of a country, equal to the value of all final goods and services produced by the factors of production supplied by the country's residents regardless of where the factors are located; GNI = GDP plus income from abroad minus income sent abroad. May be contrasted with *gross domestic product* (GDP).

growth in production possibilities An outward shift in the *PPC* caused by a decrease in unemployment, or improvement in efficiency of production, or both, leading to more output produced.

growth maximization One possible goal of firms, involving the objective to make the growth of the firm as high as possible.

growth See *economic growth*.

Happiness Index An alternative method to standard national income accounting that measures economic well-being using numerous quality of life dimensions in addition to real GDP *per capita*.

Happy Planet Index An alternative method to standard national income accounting that takes into account environmental sustainability and inequalities. It is a measure of sustainable well-being based on four dimensions, life expectancy, well-being, inequality of outcomes, ecological footprint.

homogeneous product A product that is completely standardised and not differentiated; is characteristic of products in perfect competition.

household A consumer or group of consumers (such as a family) that buys goods and services for its own use; a key decision-maker in a market economy.

household indebtedness The degree to which households have debts.

how to produce One of the three basic economic questions; refers to the choice that must be made in response to the question of what combinations of resources and what types of technologies to use in order to produce goods and services.

human capital The skills, abilities and knowledge acquired by people, as well as good levels of health, all of which make them more productive; considered to be a kind of 'capital' because it provides a stream of future benefits by increasing the amount of output that can be produced in the future.

Human Development Index (HDI) A composite indicator of development which includes indicators that measure three dimensions of development: income *per capita*, levels of health and educational attainment; is considered to be a better indicator of development than single indicators such as GNI *per capita*.

humanitarian aid Foreign aid extended in regions where there are emergencies caused by violent conflicts or natural disasters such as floods, earthquakes and tsunamis, intended to save lives, ensure access to basic necessities such as food, water, shelter and health care, and provide assistance with reconstruction.

hypothesis An educated guess, usually about a cause-and-effect relationship about an event, usually used to make predictions of real-world events by use of empirical evidence; hypotheses that have not been refuted (rejected) by the evidence may be used to build theories.

imperfect competition Refers to the situation where firms face some degree of competition (therefore they are not monopolies) but also have some degree of market power (therefore they are not perfectly competitive); include monopolistic competition and oligopoly.

import quota See *quota*.

import substitution A growth and trade strategy where a country begins to manufacture simple consumer goods oriented towards the domestic market (such as shoes, textiles, beverages, electrical appliances) in order to promote its domestic industry; it presupposes the imposition of protective measures (tariffs, quotas, etc.) that will prevent the entry of imports that compete with domestic producers. To be contrasted with *export promotion.*

imports Goods or services produced in other countries that are bought and brought into the domestic economy.

incentive function of price See *prices as incentives*.

incentive-related policies Policies involving reduction of various types of taxes (such as income taxes and business taxes), in the expectation that the tax cuts will change the incentives faced by tax-payers; for example, cuts in income taxes may encourage the desire to work; cuts in business taxes may encourage investment. Are a type of *market based supply-side policy*.

income (i) the money people receive from their employment as well as other sources including interest from savings accounts and holdings of bonds, rents from property, pensions, etc; (ii) in the current account of the balance of payments, refers to inflows of all income (mainly interest and profits) earned abroad minus the same income factors that are sent abroad.

income approach A method used to measure the value of aggregate output of an economy, which adds up all income earned by the factors of production in the course of producing all goods and services within a country in a given time period. As suggested by the circular flow model, it is equivalent to measurement by the *expenditure approach* and the *output approach*.

income distribution See *distribution of income.*

income effect Part of an explanation of the law of demand; as price falls real income increases causing the consumer to buy more of the good (the other part is the *substitution effect*).

income elastic demand Relatively high responsiveness of demand to changes in income; *YED* (income elasticity of demand) > 1. See *income elasticity of demand.*

income elasticity of demand A measure of the responsiveness of demand to changes in income; measured by the percentage change in quantity demanded divided by the percentage change in price.

income inelastic demand Relatively low responsiveness of demand to changes in income; *YED* (income elasticity of demand) < 1. See *income elasticity of demand.*

income redistribution See *redistribution of income*.

income taxes Taxes paid by households on their incomes

indicator A measurable variable that indicates the state or level of something; for example GDP per capita is an indicator of the level of output per person in an economy.

indirect taxes Taxes levied on spending to buy goods and services, called indirect because, whereas payment of some or all of the tax by the consumer is involved, they are paid to the government authorities by the suppliers (firms), that is, indirectly.

industrial policies Interventionist supply-side policies designed to support the growth of the industrial sector of an economy; may include support for small- and medium-sized firms or support for 'infant industries' through tax cuts, grants, low interest loans and other measures, as well as investment in human capital, research and development, or infrastructure development in support of industry.

inelastic Refers to a low responsiveness of a variable (such as quantity demanded) to a change in another variable (such as price or income); see various elasticities.

inequality (unequal) The absence of *equality*.

Inequality-adjusted Human Development Index (IHDI) A composite indicator that measures human development in the same three dimensions as the *Human Development Index* (income *per capita*, levels of health and educational attainment), but each dimension is adjusted for inequality in the corresponding dimension; attempts to measure losses in human development that arise from inequality.

inequity (inequitable) The absence of *equity*, when something is not equitable or fair

infant industry A new domestic industry that has not had time to establish itself and achieve efficiencies in production, and may therefore be unable to compete with more 'mature' competitor firms from abroad; considered to be one of the strongest arguments in favour of trade protection in developing countries. Also known as sunrise industry.

inferior good A good the demand for which varies negatively (or indirectly) with income; this means that as income increases, the demand for the good decreases.

inflation A continuing (or sustained) increase in the general price level.

inflation targeting A type of monetary policy carried out by some central banks that focuses on achieving a particular inflation target, rather than focusing on the goals of low and stable rate of inflation and low unemployment.

inflationary gap A situation where real GDP is greater than potential GDP, and unemployment is lower than the natural rate of unemployment; it arises when the *AD* curve intersects the *SRAS* curve at a higher level of real GDP than potential GDP.

informal economy That part of an economy that lies outside of the formal economy, consisting of economic activities that are unregistered and legally unregulated; it exists everywhere in the world but in developing countries it is often a very important part of the economy, because it offers work and income to substantial portions of the population who have no other opportunities to work in view of lack of employment opportunities in the formal economy.

infrastructure Numerous types of physical capital resulting from investments, making major contributions to economic growth and development by lowering costs of production and increasing productivity; include power, telecommunications, piped water supplies, sanitation, roads, major dam and canal works for irrigation and drainage, urban transport, ports and airports.

injections In the circular flow of income model, refer to the entry into income flow of funds corresponding to investment, government spending or exports.

integration See *economic integration*.

interdependence One of the key concepts of this course; refers to the idea that economic decision-makers interact with and depend on each other; arises from the fact that no one is self-sufficient. *One of the key concepts of this course*

interest (i) A payment, per unit of time, for the use of borrowed money (borrowers pay interest, lenders receive interest). (ii) A payment, per unit of time, to owners of capital resources.

interest rate Interest expressed as a percentage; in the case of borrowed money, it is interest as a percentage of the amount borrowed. Changes in interest rates form the basis of *monetary policy*.

International Monetary Fund (IMF) An international financial institution composed of 189 member countries, whose purpose is to make short-term loans to governments on commercial terms (i.e. non-concessional) in order to stabilise exchange rates, alleviate balance of payments difficulties and help countries meet their foreign debt obligations.

intervention One of the key concepts of this course; typically refers to government intervention, meaning that the government becomes involved with the workings of markets..

interventionist policy Any policy based on government intervention in the market; to be contrasted with *market-oriented policy*. See also *government intervention*.

interventionist supply-side policy Any policy based on government intervention in the market intended to affect the supply-side of the economy, usually to shift the *LRAS* curve to the right, increase potential output and achieve long term economic growth; see industrial policy or investments in education or infrastructure as examples. May be contrasted with market-based supply side policy.

investment Includes spending by firms or the government on capital goods (i.e. buildings, machinery, equipment, etc.) and all spending on new construction (housing and other buildings).

J-curve effect A curve that plots the balance of trade (exports minus imports) on the vertical axis and time on the horizontal axis, showing that a country with a devaluing/depreciating currency may see a worsening in its trade balance (an increase in a trade deficit) in the period immediately following the devaluation or depreciation, while in a later period the trade deficit will begin to shrink provided the Marshall–Lerner condition holds (see *Marshall–Lerner condition*).

John Maynard Keynes (1883–1946) A British economist; one of his ideas that he is famous for is that an economy left on its own will not necessarily lead to full employment on its own, thus requiring government intervention.

joint supply Refers to production of two or more goods that are derived from a single product, so that it is not possible to produce more of one without producing more of the other (for example, butter and skimmed milk are both produced from whole milk, and producing more of one means producing more of the other as well).

Karl Marx A German philosopher who developed a theory predicting the collapse of capitalism and its replacement by communism, highly regarded for his insights into how capitalism works.

Keynesian aggregate supply curve An aggregate supply curve that has a flat (horizontal) section, an upward sloping section and a vertical section. It shows the relationship between real GDP and the price level on the assumption that prices and wages are inflexible downward. Changes in the price level and/or real GDP depend on the level of aggregate demand and where the economy is producing relative to full capacity output.

Keynesian multiplier The ratio of real GDP divided by a change in any of the components of aggregate spending (consumption C, investment I, government spending G, or net exports $X - M$); alternatively it is $1/(1-MPC)$, where MPC is the marginal propensity to consume. The value of this ratio is usually greater than one because of a multiplied effect of an initial change in a component of aggregate spending on the final value of real output.

Keynesian revolution A school of thought based on the contributions of John Maynard Keynes, according to which government intervention in the economy is essential for the management of aggregate demand in order to ensure that full employment will be achieved.

labour A factor of production, which includes the physical and mental effort that people contribute to the production of goods and services.

labour market flexibility Refers to the operation of market forces (supply and demand) in the labour market. May be achieved by reducing or eliminating interference with market forces (for example, reducing or eliminating minimum wages and labour union activities, reducing job security, etc.).

labour union an association of workers in a particular profession, whose objective is to improve working conditions and defend rights of workers, representing its members in negotiations with employers.

laissez-faire A French expression meaning *let it do*, referring to a free market economy without government intervention, such as was advocated by *Adam Smith*. According to this idea a free market economy left on its own will be highly efficient.

land A factor of production which includes all natural resources: land and agricultural land, as well as everything that is under or above the land, such as minerals, oil reserves, underground water, forests, rivers and lakes. Natural resources are also called 'gifts of nature' or 'natural capital'.

land rights A group of property rights that refer to the rights and rules to possess, occupy and use land; the term *land* here refers to natural resources including the land itself, trees, minerals, pasture, water.

law of demand A law stating that there is a negative relationship between the price of a good and quantity of the good demanded, over a particular time period, *ceteris paribus*: as the price of the good increases, the quantity of the good demanded falls (and vice versa).

law of diminishing marginal returns A law stating that as more and more units of a variable input (such as labour) are added to one or more fixed inputs (such as land), the marginal product of the variable input at first increases, but there comes a point when the marginal product of the variable input begins to decrease; underlies the firms' supply curve.

law of diminishing marginal utility A law stating that as consumption of a good increases, marginal utility, or the extra utility the consumer receives, decreases with each additional unit consumed, therefore consumers will buy more only if price falls; this underlies the law of demand.

law of supply A law stating that there is a positive relationship between the price of a good and quantity of the good supplied, over a particular time period, *ceteris paribus*: as the price of the good increases, the quantity of the good supplied also increases (and vice versa).

leakages In the circular flow of income model, refers to the withdrawal from the income flow of funds corresponding to savings, taxes or imports.

logic A method of reasoning which involves making a series of statements each of which is true if the preceding statement is true.

long run (in macroeconomics) In macroeconomics, it is the period of time when prices of resources (especially wages) change along with changes in the price level.

long run (in microeconomics) In microeconomics, it is a time period in which all inputs can be changed; there are no fixed inputs.

long-run aggregate supply (*LRAS*) curve A curve showing the relationship between real GDP produced and the price level when wages (and other resource prices) change to reflect changes in the price level, *ceteris paribus*. The *LRAS* curve is vertical at the full employment level of GDP, indicating that in the long run output is independent of the price level.

long-run aggregate supply The total quantity of goods and services (real output or real GDP) produced in an economy in the *long run* (when wages and other resource prices change to reflect changes in the price level), *ceteris paribus*.

long-run equilibrium level of output The level of output (real GDP) that results when the economy is in long-run equilibrium, occurring when the aggregate demand and short-run aggregate supply curves intersect at a point on the long-run aggregate supply curve; occurs where the vertical *LRAS* curve intersects the horizontal axis, known as *potential output*.

long-run Phillips curve See *Phillips curve*.

long-term growth Growth of an economy (growth in real output) over long periods of time; shown by rightward shifts of the long-run aggregate supply (LRAS) curve corresponding to the long-term growth trend of the business cycle; or outward shifts of the production possibilities curve (PPC).

long-term growth trend In the business cycle diagram, refers to the line that runs through the business cycle curve, representing average growth over long periods of time; shows how output grows over time when cyclical fluctuations are ironed out. The output represented by the long-term growth trend is known as *potential output*.

Lorenz curve A curve illustrating the degree of equality (or inequality) of income (or wealth) distribution in an economy. It plots the cumulative percentage of income received by cumulative shares of the population. The closer the Lorenz curve is to the diagonal line of perfect equality, the more equal the income distribution.

loss Arises when total costs of a firm are greater than its total revenue; alternatively when average cost is greater than average revenue, or average cost is greater than price; it is negative *profit*.

luxury goods Goods that are not necessary or essential; they have a *price elastic demand* (PED>1) and *income elastic demand* (YED>1). To be contrasted with *necessities*.

macroeconomic equilibrium In macroeconomics occurs where aggregate demand interests aggregate supply, determining the price level and level of real output (real GDP); see *short-run* and *long-run equilibrium level of output*.

macroeconomics The branch of economics that examines the economy as a whole by use of aggregates, which are wholes or collections of many individual units, such as the sum of consumer behaviours and the sum of firm behaviours, total income and output of the entire economy as well as total employment and the general price level.

managed exchange rates Exchange rates that are for the most part free to float to their market levels (i.e. their equilibrium levels) over long periods of time; however, central banks periodically intervene in order to stabilise them over the short term.

mandated choice A term from *behavioural economics,* it is a choice between alternatives that is made mandatory (compulsory) by the government or other authority; required choice.

manufactured products goods produced by labour usually working together with capital as well as raw materials, such as for example cars, computers and televisions.

margin/marginal Means extra or additional; is used in a number of contexts in economics, including *marginal utility*, *marginal benefit*, *marginal cost*, *marginal product*, *marginal revenue*, *marginal tax rate*.

marginal benefit The extra or additional benefit received from consuming one more unit of a good.

marginal cost The extra or additional cost of producing one more unit of output.

marginal private benefits (*MPB*) The extra or additional benefit received by consumers when they consume one more unit of a good.

marginal private costs (*MPC*) The extra or additional costs to producers of producing one more unit of a good.

marginal product The extra or additional output that results from one additional unit of a variable input (such as labour).

marginal propensity to consume (*MPC*) The fraction of additional income spent on domestically produced goods and services. Determines the size of the Keynesian multiplier; the larger the *MPC*, the larger the multiplier.

marginal propensity to import (*MPM*) The fraction of additional income spent on imports. The larger the *MPM*, the smaller the Keynesian multiplier.

marginal propensity to save (*MPS*) The fraction of additional income that is saved. The larger the *MPS*, the smaller the Keynesian multiplier.

marginal propensity to tax (*MPT*) The fraction of additional income that is paid as taxes. The larger the *MPT*, the smaller the Keynesian multiplier.

marginal revenue The additional revenue arising from the sale of an additional unit of output.

marginal social benefits (*MSB*) The extra or additional benefits to society of consuming one more unit of a good; are equal to marginal private benefits (*MPB*) when there are no consumption externalities.

marginal social costs (*MSC*) The extra or additional costs to society of producing one more unit of a good; are equal to marginal private costs (*MPC*) when there are no production externalities.

marginal tax rate The tax rate paid on additional income; refers to the tax rate that applies to the highest tax bracket of an individual's personal income.

marginal utility (MU) The extra or additional utility received from consuming one more unit of a good.

market Any kind of arrangement where buyers and sellers of a particular good, service or resource are linked together to carry out an exchange.

market concentration The degree to which a market is dominated by a small number of large firms. The smaller the number of firms controlling a market, the greater the market concentration. It is measured by the *concentration ratio*.

market demand Refers to the sum of all individual consumer demands for a good or service.

market equilibrium Occurs where quantity demanded is equal to quantity supplied, and there is no tendency for the price or quantity to change.

market failure Occurs when the market fails to allocate resources efficiently, or to provide the quantity and combination of goods and services mostly wanted by society. Market failure results in allocative inefficiency, where too much or too little of goods or services are produced and consumed from the point of view of what is socially most desirable.

market power Refers to the control that a seller may have over the price of the product it sells; it is the ability of a firm to charge a price greater than marginal cost, or $P>MC$.

market share Refers to the percentage of total sales in a market that is earned by a single firm; it may be a goal of some firms to maximise market share, or to make it as large as possible.

market supply Refers to the sum of all individual firm supplies of a good or service.

market-based supply-side policy Any policy based on promoting well-functioning, competitive markets in order to influence the supply-side of the economy, usually to shift the *LRAS* curve to the right, increase potential output and achieve long term economic growth; include labour market reforms, competition policies and incentive-related policies. May be contrasted with *interventionist supply-side policy*.

market-oriented policy A policy in which government intervention is limited, economic decisions are made mainly by the private decision-makers (firms and consumers) and the market has significant freedom to determine resource allocation; to be contrasted with *interventionist policy*.

Marshall–Lerner condition A condition stating when depreciation or devaluation of a country's currency will lead to an improvement in that country's balance of trade: the sum of the price elasticities of demand for imports and exports must be greater than 1 for the trade balance to improve (for a trade deficit to become smaller). This usually holds over the longer term, but not in the shorter term (see *J-curve*).

maximum price See *price ceiling*.

merit goods Goods that are held to be desirable for consumers, but which are underprovided by the market. Reasons for underprovision are usually that the goods have *positive consumption externalities*; in addition, consumers with low incomes may be unable afford them (and so do not demand them), or there is consumer ignorance about the benefits of the goods.

microeconomics The branch of economics that examines the behaviour of individual decision-making units, consumers and firms; is concerned with consumer and firm behaviour and how their interactions in markets determine prices in goods markets and resource markets.

microfinance Programmes to provide credit (loans) in small amounts to people who do not ordinarily have access to credit due to lack of collateral; also known as microcredit. These are very important in developing countries as a strategy to help poor people climb out of poverty.

minimum income standards A method to measure poverty consisting of ongoing research on what people in a population believe are the essentials for a minimum acceptable standard of living that allows people to participate in society; is used to create a basket of goods needed to achieve this minimum.

minimum lending rate The interest rate charged by the central bank when it lends funds to commercial banks (according to UK terminology); has different names in different countries, for example *refinancing rate* in the European Union, *discount rate* in the United States, *base rate* in the United Kingdom.

minimum price See *price floor*.

minimum reserve requirement A legally determined fraction of total deposits of commercial banks that must be kept within their vaults, which cannot be lent out.

minimum wage A minimum price of labour (the 'wage') set by governments in the labour market, in order to ensure that low-skilled workers can earn a wage high enough to secure them with access to basic goods and services. It is a type of *price floor*.

mixed economy An economy that combines the command approach (government decision-making) with the market approach (private sector decision-making) to resource ownership, decision-making and rationing.

mobile banking Refers to the use of mobile phones and the internet to pay bills and to receive or pay money, meeting the needs of users for easier banking transactions; is playing an important role in reducing poverty in developing countries.

model A simplified representation of something in the real world, showing only the important aspects of the real world being investigated, ignoring unnecessary details.

monetarist/new classical model Actually includes two different models of the macroeconomy (the monetarist and the new classical); both are based on the following principles: the importance of the price mechanism in co-ordinating economic activities, the concept of competitive market equilibrium, and thinking about the economy as a harmonious system that automatically tends toward full employment.

monetary policy Policy carried out by the central bank, aiming to change interest rates in order to influence aggregate demand; it is a type of *demand-side policy*, or *demand management*.

monetary union A high form of economic integration, involving the adoption by a group of countries of a single currency, such as some of the countries of the European Union ('euro zone' countries) that have adopted the euro. Monetary union in addition involves the adoption of a common monetary policy carried out by a single central bank, which is necessitated by the use of a single currency.

money Anything that is acceptable as payment for goods and services; more precisely, money consists of currency (coins and paper money) and checking accounts.

money creation Refers to the process by which commercial banks create new money when they make loans.

monopolistic competition One of the four market structures, with the following characteristics: a large number of firms; substantial control over market price; product differentiation; no barriers to entry. Examples include the shoe, clothing, detergent, computer, publishing, furniture and restaurant industries.

monopoly One of the four market structures, with the following characteristics: a single or dominant large firm in the industry; significant control over price; produces and sells a unique product with no close substitutes; high barriers to entry into the industry. Examples include telephone, water and electricity companies in areas where they operate as a single supplier (these are *natural monopolies*).

moral hazard Refers to situations where one party takes risks, but does not face the full costs of these risks because the full costs of the risks are borne by the other party; often arises in connection with insurance; is a form of *asymmetric information* causing market failure.

Multidimensional Poverty Index (MPI) A composite indicator that measures poverty in three dimensions: health, education and living standards, each of which reflects deprivations (essential things people don't have).

multilateral development assistance Lending to developing countries for the purpose of assisting their development on non-concessional terms (market rates of interest and repayment periods) by multilateral organisations, i.e. organisations composed of many countries, including development banks such as the World Bank, and the International Monetary Fund; to be contrasted with *foreign aid*.

multilateral trade agreement A trade agreement (or agreement to lower international trade barriers) between many countries; at the present time these are mainly carried out within the framework of the *World Trade Organization (WTO)*, and involve agreements between WTO member countries. May be contrasted with *bilateral trade agreement* and *regional trade agreement*.

multinational corporation (MNC) A firm involved in *foreign direct investment (FDI)*; it is a firm that is based in one country (the home country) and that undertakes productive investments in another country (the host country) with control of at least 10 per cent of the firm in the host country.

multiplier See *Keynesian multiplier*.

national debt See *government debt*

national income accounting Measurement of an economy's national income and output as well as other measures of economic performance by specialised statistical services in every country.

national income statistics Statistical data used to measure national income and output and other measures of economic performance.

national income The total income of an economy, consisting of factor payments or the sum of wages, interest, rent plus profit, often used interchangeably with the value of aggregate output, particularly in the context of macroeconomic models (such as the *AD-AS* model).

national output Total output produced by an economy, also known as aggregate output, often measured by *real GDP*.

natural monopoly A single firm (a monopoly) that can produce for the entire market at a lower average cost than two or more smaller firms. This happens when the market demand for the monopolist's product is within the range of falling long-run average cost, where there are *economies of scale*. Examples include telephone, water and electricity companies in areas where they operate as a single supplier.

natural rate of unemployment Unemployment that occurs when the economy is producing at its potential or full employment level of output (real GDP), and is equal to the sum of structural, frictional plus seasonal unemployment.

necessities Goods that are necessary or essential: they have a *price inelastic demand* (PED<1) and *income inelastic demand* (YED<1). To be contrasted with *luxuries*.

negative consumption externality A negative externality caused by consumption activities, leading to a situation where marginal social benefits are less than marginal private benefits ($MSB < MPB$); see also *externality* and *negative externality.*

negative externality A type of externality where the side-effects on third parties are negative or harmful. To be contrasted with *positive externality*; see also *externality*.

negative production externality A negative externality caused by production activities, leading to a situation where marginal social costs are greater than marginal private costs ($MSC > MPC$); see also *externality* and *negative externality*.

net exports Refers to the value of exports minus the value of imports.

nominal GDP Gross domestic product measured in terms of current (or nominal) prices, which are prices prevailing at the time of measurement. Does not account for changes in the price level; to be distinguished from *real GDP*.

nominal GNI Gross product national income in terms of current (or nominal) prices, which are prices prevailing at the time of measurement. Does not account for changes in the price level; to be distinguished from *real GNI*

nominal interest rate An interest rate that prevails at any moment in time, which does not take changes in the price level into account; to be distinguished from *real interest rate*.

nominal value Value that is in money terms, measured in terms of prices that prevail at the time of measurement, and that does not account for changes in the price level; to be distinguished from *real* values.

non-collusive oligopoly A type of oligopoly where firms do not make agreements among themselves (i.e. do not collude) in order to fix prices or collaborate in some way.

non-excludable A characteristic of some goods where it is not possible to exclude someone from using a good, because it is not possible to charge a price; it is one of the two characteristics of public goods (to be contrasted with *excludable*). See also *non-rivalrous*.

non-governmental organisations (NGOs) Non-profit organisations that provide a very wide range of services and humanitarian functions; in developing countries they provide foreign aid, all of which takes the form of grants (there are no loans involved). They are involved with an enormous range of activities, including emergency assistance, promotion of sustainable development, poverty alleviation, protection of child health, provision of technical assistance, and many more.

non-price competition Occurs when firms compete with each other on the basis of methods other than price (such as product differentiation, advertising and branding). Non-price competition occurs in oligopoly and monopolistic competition.

non-price determinants of demand The variables (other than price) that can influence demand, and that determine the position of a demand curve; a change in any determinant of demand causes a shift of the demand curve.

non-price determinants of supply The variables (other than price) that can influence supply, and that determine the position of a supply curve; a change in any determinant of supply causes a shift of the supply curve.

non-produced, non-financial assets A part of the capital account of the balance of payments, which includes a variety of items such as mineral rights, forestry rights, fishing rights and airspace.

non-rivalrous A characteristic of some goods where the consumption of the good by one person does not reduce consumption by someone else; it is one of the two characteristics of public goods (to be contrasted with *rivalrous*). See also *non-excludable*.

normal good A good the demand for which varies positively (or directly) with income; this means that as income increases, demand for the good increases.

normal profit The minimum amount of revenue that a firm must receive so that it keeps the business running (as opposed to shutting down). Normal profit is earned when abnormal profit is zero, or when average revenue is equal to average cost, or price is equal to average cost.

normative economics The body of economics based on normative statements, which involve beliefs, or value judgements about what ought to be. Normative statements cannot be true or false; they cannot be refuted, they can only be assessed relative to beliefs and value judgements. Normative economics forms the basis of economic policies; to be contrasted with *positive economics*.

nudge Part of *behavioural economics,* it is a method designed to influence consumers' choices in a predictable way, without offering financial incentives or imposing sanctions, and without limiting choice.

OECD Better Life index An alternative measure to standard national income accounting that measures economic well-being in a number of dimensions that take into account quality of life.

official borrowing Refers to government borrowing from abroad; is an item in the *financial account* of a country's balance of payments.

Official Development Assistance (ODA) The most important part of *foreign aid*, referring to foreign aid that is offered by countries or by international organisations composed of a number of countries (it does not include aid offered by non-governmental organisations).

oligopoly One of the four market structures, with the following characteristics: small number of large firms in the industry; firms have significant control over price; firms are interdependent; products may be differentiated or homogeneous; there are high barriers to entry. Examples include the car industry, airlines, electrical appliances (differentiated products) and the steel, aluminium, copper, cement industries (homogeneous products).

open market operations A tool of monetary policy whereby the central bank buys and sells bonds to commercial banks in order to influence the money supply and interest rate.

opportunity A set of circumstances making it possible for someone to do something.

opportunity cost The value of the next best alternative that must be given up or sacrificed in order to obtain something else.

output approach A method used to measure the value of aggregate output of an economy, which calculates the value of all final goods and services produced in the country within a given time period. As suggested by the circular flow model, it is equivalent to measurement by the *expenditure approach* and the *income approach.*

overvalued currency A currency whose value is higher than its free-market value; may occur if the exchange rate is fixed (or pegged), or in a managed exchange rate system, but not in a freely floating exchange rate system. To be contrasted with *undervalued currency*.

payoff matrix In game theory, this shows all possible combinations of outcomes of different decisions made by the players in game theory

per capita Per person, or per head. For example, GDP *per capita* is total GDP divided by the number of people in the population.

perfect competition One of the four market structures, with the following characteristics: a large number of small firms; no control over price; all firms sell a homogeneous product; no barriers to entry, perfect information and perfect resource mobility. Close examples include agricultural commodity markets and the foreign exchange market.

perfectly elastic demand Refers to a price elasticity of demand value of infinity, and arises in the case of a horizontal demand curve indicating that any quantity can be bought at that price; see *price elasticity of demand*.

perfectly elastic supply Refers to a price elasticity of supply value of infinity, and arises in the case of a horizontal supply curve indicating that any amount can be sold at that price; see *price elasticity of supply*.

perfectly inelastic demand Refers to a price elasticity of demand value of zero, and arises in the case of a vertical demand curve indicating that any amount can be sold at that price; see *price elasticity of demand*.

perfectly inelastic supply Refers to a price elasticity of supply value of zero, and arises in the case of a vertical supply curve; see *price elasticity of supply*.

personal income taxes Taxes paid by households or individuals in households on all forms of income, including wages, rental income, interest income, and dividends (income from ownership of shares in a company).

Phillips curve A curve showing the relationship between unemployment and inflation. The short-run Phillips curve shows a negative relationship between the rate of inflation and the unemployment rate. The long-run Phillips curve is vertical at the natural rate of unemployment, indicating that there is no negative relationship between inflation and unemployment.

Pigouvian taxes (or Pigovian taxes) *Indirect taxes* designed to correct negative externalities of production or consumption.

planned economy An economy where all economic decision-making is carried out by government planning (based on *command and control* methods); rather than reliance on prices determined in markets; to be contrasted with a *free market economy*.

portfolio investment Financial investment, including investment in stocks and bonds. Appears as an item in the *financial account* of the balance of payments.

positive consumption externality A positive externality caused by consumption activities, leading to a situation where marginal social benefits are greater than marginal private benefits ($MSB > MPB$); see also *externality* and *positive externality.*

positive economics The body of economics based on positive statements, which are about things that are, were or will be. Positive statements may be true or false so they can be refuted. They form the basis of theories and models that try to explain economic events. To be contrasted with *normative economics*.

positive externality A type of externality where the side-effects on third parties are positive or beneficial, also known as 'spillover benefits'; to be contrasted with *negative externality*; see also *externality*.

positive production externality A positive externality caused by production activities, leading to a situation where marginal social costs are less than marginal private costs ($MSC < MPC$); see also *externality* and *positive externality.*

potential output (potential GDP) The level of output (real GDP) that can be produced when there is 'full employment', meaning that unemployment is equal to the natural rate of unemployment; also known as the *full employment level of output*.

poverty The inability of an individual or family to afford an adequate standard of goods and services; this standard may be absolute or relative; see *absolute poverty* and *relative poverty*.

poverty cycle (trap) Arises when low incomes result in low (or zero) savings, permitting only low (or zero) investments in physical, human and natural capital, and therefore low productivity of labour and of land, which in turn gives rise to low, if any, growth in income (sometimes growth may be negative), and hence low incomes once again. Poverty is transmitted from generation to generation.

poverty line An income level that is just enough to ensure a family with the minimum necessary food, housing, clothing, medical needs.

poverty trap See *poverty cycle*.

preferential trade agreement An agreement between two or more countries to lower trade barriers between them on particular products, resulting in easier access to the markets of other members for the selected products, compared with the access of countries that are not members.

price ceiling A maximum price set by the government for a particular good, meaning that the price that can be legally charged by the sellers of the good cannot be higher than the legal maximum price. Results in a *shortage* of the product.

price competition Occurs when a firm lowers its price to attract customers away from rival firms, thus increasing sales at the expense of other firms. May occur in the case of monopolistic competition or oligopoly, but not in perfect competition (or monopoly).

price control Setting of minimum or maximum prices by the government (or private organisations) so that prices are unable to adjust to their equilibrium level determined by demand and supply. Price controls result in shortages or surpluses.

price deflator A price index used to convert nominal values into real values, such as nominal GDP into real GDP, known as the 'GDP deflator'

price elastic demand Relatively high responsiveness of quantity demanded to changes in price; *PED* (price elasticity of demand) > 1. See *price elasticity of demand.*

price elastic supply Relatively high responsiveness of quantity supplied to changes in price; *PES* (price elasticity of supply) > 1. See *price elasticity of supply*.

price elasticity of demand (*PED*) A measure of the responsiveness of the quantity of a good demanded to changes in its price, given by the percentage change in quantity demanded divided by the percentage change in price. In general, if there is a large responsiveness of quantity demanded ($PED > 1$), demand is referred to as being elastic; if there is a small responsiveness ($PED < 1$), demand is inelastic.

price elasticity of supply (*PES*) A measure of the responsiveness of the quantity of a good supplied to changes in its price, given by the percentage change in quantity supplied divided by the percentage change in price. In general, if there is a large responsiveness of quantity supplied ($PES > 1$), supply is referred to as being elastic; if there is a small responsiveness ($PES < 1$), supply is inelastic.

price floor A minimum price set by the government for a particular good, meaning that the price that can be legally charged by the sellers of the good cannot be lower than the legal minimum price. Results in a *surplus* of the product.

price inelastic demand Relatively low responsiveness of demand to changes in price; *PED* (price elasticity of demand) < 1. See *price elasticity of demand*.

price inelastic supply Relatively low responsiveness of supply to changes in price; *PES* (price elasticity of supply) < 1. See *price elasticity of supply*.

price mechanism The system where prices are determined by demand and supply in competitive markets, resulting from the free interaction of buyers (demanders) and sellers (suppliers); these interactions determine the allocation of resources.

price war Competitive price-cutting by firms; usually in oligopoly. As each one tries to capture market shares from rival firms; results in lower profits for firms.

price-maker Any firm that has the ability to influence the price of its product; arises whenever the firm faces a downward slopping demand curve; applies to varying degrees in all market structures except perfect competition.

price-taker A firm that accepts a price at which it sells its product. Refers to firms in perfect competition, which being small and numerous have no control over price, and therefore accept the price determined in the market.

prices as incentives The ability of prices, and changes in prices, to convey information to consumers and producers that motivates them to respond by offering them incentives to behave in their best-self-interest; firms according to the law of supply and consumers according to the law of demand; compare with *prices as signals*, which together with prices as incentives lead to an efficient allocation of resources (assuming no market failures).

prices as signals The ability of prices, and changes in prices, to communicate information to consumers and producers about the existence of excess demand or excess supply, on the basis of which they make economic decisions, which together with prices as incentives lead to an efficient allocation of resources (assuming no market failures).

primary commodity Any product that is produced in the *primary sector*, which includes agriculture, forestry, fishing and the extractive industries; also known as 'commodity'.

primary products All products produced in the primary sector of an economy; also known as commodities; see *primary sector*.

primary sector A part of an economy that is dominated by agriculture, also including fishing, forestry and all extractive activities (such as mining).

privatisation A transfer of ownership from the public sector (the government) to the private sector, i.e. private owners.

producer surplus Refers to the difference between the price received by firms for selling their good and the lowest price they are willing to accept to produce the good. In a diagram, it is shown as the area under the price received by producers and above the supply curve up to the quantity sold.

product differentiation Occurs when each firm in an industry tries to make its product different from those of its competitors; usually in order to create some market power (monopoly power); products can be differentiated by physical differences, quality differences, location, services and product image.

production possibilities All possible combinations of the maximum amounts of two goods that can produced by an economy, given fixed and unchanging resources and technology, when there is full employment of resources and efficiency in production.

production possibilities curve (*PPC*) A curve showing *production possibilities*.

profit A payment, per unit of time, to owners of entrepreneurship/management (a factor of production). See *abnormal profit* and *normal profit*.

profit maximisation The goal of firms, according to the standard theory of the firm. It involves making profit as large as possible, and is achieved by producing the level of output where the difference between total revenue and total costs is the largest, or where marginal cost is equal to marginal revenue.

progressive taxation Taxation where, as income increases, the fraction of income paid as taxes increases; there is an increasing average tax rate.

property rights The laws and regulations that define rights to ownership, use and transfer of property.

proportional taxation Taxation where, as income increases, the fraction of income paid as taxes remains constant; there is a constant average tax rate.

public good A good that is *non-rivalrous* (its consumption by one person does not reduce consumption by someone else) and *non-excludable* (it is not possible to exclude someone from using the good). Since it is not possible to exclude someone from using the good even though they do not pay for it, firms do not have an incentive to produce it. Public goods are therefore provided by the government. This is a type of market failure.

purchasing power parity (PPP) exchange rates Special exchange rates between currencies that makes the buying power of each currency equal to the buying power of US$1, and therefore equal to each other. The use of PPP exchange rates to convert GDP (or GNI or any other output or income variable) eliminates the influence of price level differences across countries and is very important for making cross-country comparisons.

quantitative easing A tool used by central banks to increase the money supply in the economy and facilitate commercial bank lending as part of expansionary monetary policy; involves the buying of bonds by the central bank on a large scale.

quintiles Division of a population into five equal groups with respect to the distribution of a variable, such as income; for example, the lowest income quintile refers to 20% of the population with the lowest income.

quota A type of trade protection that involves setting a legal limit to the quantity of a good that can be imported over a particular time period (typically a year). (More generally, a 'quota' is a limited or fixed number of things.)

rate of interest See *interest rate*.

rational consumer choice In the microeconomic theory of consumer behavior, consumers make choices about what goods and services to buy based on the following assumptions. All consumers (i) have consistent tastes and preferences, (ii) have perfect information, and (iii) try to maximise their *utility* (make it as great as possible).

rational producer behaviour The basis of standard theory of the firm according to which firms try to maximise profit.

rationing A method used to apportion or divide something up between its interested users; in economics it refers to the method used to make resource allocation and output/income distribution decisions.

real GDP Gross domestic product (GDP) measured in constant prices, i.e. prices that prevail in one particular year, called a 'base year'; this is useful for making comparisons of changes in GDP over time that have taken into account the influence of changing prices; to be distinguished from nominal GDP.

real GNI Gross national income (GNI) measured in constant prices, i.e. prices that prevail in one particular year, called a 'base year'; this is useful for making comparisons of changes in GNI over time that have taken into account the influence of changing prices; to be distinguished from nominal GNI.

real interest rate The interest rate that has been corrected for inflation; real interest rate = nominal interest rate – rate of inflation.

real value Value that has eliminated the influence of changes in the price level.

recession An economic contraction, where there is falling real GDP (negative growth) and increasing unemployment of resources which last six months or more.

recessionary gap A situation where real GDP is less than potential GDP, and unemployment is greater than the natural rate of unemployment; it arises when the *AD* curve intersects the *SRAS* curve at a lower level of real GDP than potential GDP. Also known as 'deflationary gap'.

redistribution of income Refers to changing the distribution of income, giving rise to a new distribution.

refinancing rate See *minimum lending rate*.

refutation In the sciences and social sciences it is the idea that it must be possible to *refute or disprove* a hypothesis or a theory by subjecting it to empirical testing.

regional trade agreement A trade agreement (or agreement to lower international trade barriers) between several countries that are located within a geographical region (such as NAFTA, or North American Free Trade Agreement). May be contrasted with *bilateral trade agreement and multilateral trade agreement*.

regressive taxation Taxation where, as income increases, the fraction of income paid as taxes decreases; there is a decreasing average tax rate.

relative poverty The inability of an individual or a family to afford an adequate standard of goods and services, where the adequate standard is relative and changes over time; this standard is defined as what is 'typical' in a society, taken to be a particular percentage (often 50%) of society's median income. As incomes increase and the median income rises, the standard also rises.

remittances A transfer of money from one country to another, in most cases by foreign workers who send money from their earnings in the country of residence to their family in their home country.

reserve assets Refers to foreign currency reserves that the central bank maintains and can buy or sell to influence the value of the country's currency exchange rate; in the balance of payments appears as an item in the *financial account*.

resource allocation Assigning available resources, or factors or production, to specific uses chosen among many possible and competing alternatives; involves answering the *'what to produce'* and *'how to produce'* basic economic questions.

resources Factors of production, used by firms as inputs in the production process; see *factors of production*.

restricted choice A term in *behavioural economics*, it is a choice that is limited by the government or other authority.

revaluation (of a currency) Refers to an increase in the value of a currency in the context of a fixed or pegged exchange rate system (compare with *appreciation*, which is an increase in currency value in the contest of a floating or managed exchange rate system).

revenues The payments that firms receive when they sell their goods and services.

rivalrous A characteristic of a good according to which its consumption by one person reduces its availability for someone else; most goods are rivalrous. It is one of the two characteristics of 'private goods'. See also *excludable*.

rules of thumb A term from *behavioural economics*, are simple guidelines based on experience and common sense, simplifying complicated decisions that would have to be based on the complex consideration of every possible choice.

satisficing A goal of firms to achieve satisfactory results, rather than pursue a single maximising objective, such as to maximise profits or revenues; based on the argument that large, modern firms have numerous objectives which may partly overlap or conflict, thus forcing them to compromise and reconcile conflicts, rather than pursue optimal results.

scarcity One of the key concepts of this course; it is the condition in which available resources (land, labour, capital, entrepreneurship) are limited; they are not enough to produce everything that human beings need and want. *One of the key concepts of this course*

screening In the event of asymmetric information, this is a method used by the buyer when the buyer has limited information; for example the buyer may research the seller or product online, or may informally ask friends in order to get more information.

seasonal unemployment A type of unemployment that occurs when the demand for labour in certain industries changes on a seasonal basis because of variations in needs; for example, farm workers are hired during peak harvesting seasons and let off for the rest of the year.

short run (in macroeconomics) In macroeconomics, it is the period of time during which the prices of resources, particularly the price of labour (wages) do not change (they are constant).

short run (in microeconomics) In microeconomics, it is a time period during which at least one input is fixed and cannot be changed by the firm.

short-run aggregate supply The total quantity of goods and services (real output or real GDP) produced in an economy in the *short run* (when wages and other resource prices are held constant), *ceteris paribus*.

short-run aggregate supply (SRAS) curve A curve showing the relationship between real GDP produced and the economy's price level when wages (and other resource prices) are held constant, *ceteris paribus*; the SRAS curve is upward sloping.

short-run equilibrium level of output In the monetarist/new classical model, it is the level of output (real GDP) determined by the intersection of the aggregate demand and short run aggregate supply curves; in the Keynesian model, it is the level of output determined by the intersection of the aggregate demand and Keynesian aggregate supply curves. In both models, equilibrium may occur where there is (i) a recessionary (deflationary) gap, (ii) an inflationary gap, or (iii) full employment output.

short-run Phillips curve See *Phillips curve*.

short-term fluctuations Alternating periods of expansion (increasing real output) and contraction (decreasing real output), which occur in the *business cycle*.

short-term growth Growth of an economy (growth in real output) over relatively short periods of time; shown by (i) a movement of a point inside the production possibilities curve (PPC) to a point closer to the PPC (see *actual growth*), or (ii) by upturns in the business cycle usually due to increases in aggregate demand.

shortage In the context of demand and supply, is the amount by which quantity demanded is greater than quantity supplied.

signalling function of price See *prices as signals*.

signalling In the event of asymmetric information this is a method used by the seller when the seller has more information, which attempts to convince the buyer that the product is of good quality; for example use of warranties or establishment of brand names.

social enterprise A type of commercial organisation that aims to achieve particular social goals in an effort to improve people's well-being and promote social change; may be either for-profit or not-for-profit organisations.

socially optimum output Refers to a level of output that is the best from the socially point of view, determined by the achievement of allocative efficiency (or economic efficiency); occurs when marginal social benefits are equal to marginal social costs ($MSB = MSC$).

social science Any academic discipline that studies human society and social relationships, concerned with discovering general principles describing how societies function and are organised; include anthropology, economics, political science, psychology, sociology and others.

social surplus The sum of consumer and producer surplus; it is maximum in a competitive market with no market failures. See *consumer surplus* and *producer surplus*.

specialisation Occurs when a firm or a country concentrates production on one or a few goods and services. In international trade theory, specialisation forms the basis for the gains from trade, arising when countries specialise according to their comparative advantage, and when firms specialise in production of goods and services that offer them economies of scale.

speculation (currency) Buying and selling of something in the hope of making a profit. 'Currency speculation' involves buying and selling currencies based on expectations of changes in the value of a currency (exchange rates) in order to make a profit in the future.

structural unemployment A type of unemployment that occurs as a result of technological changes and changing patterns of demand (causing changes in demand for labour skills), as well as changes in the geographical location of jobs, and labour market rigidities (lack of *labour market flexibility*).

subsidy An amount of money paid by the government to firms for a variety of reasons: to prevent an industry from failing, to support producers' incomes, or as a form of protection against imports (due to the lower costs and lower prices that arise from the subsidy). A subsidy given to a firm results in a higher level of output and lower price for consumers. May also be paid to consumers as financial assistance or for income redistribution.

substitute goods Two or more goods that satisfy a similar need, so that one good can be used in place of another. If two goods are substitutes, an increase in the price of one leads to an increase in the demand for the other.

substitution effect Part of an explanation of the law of demand; there is an inverse relationship between price and quantity demanded because as price falls consumers substitute the now less expensive good for other products (the other part is the *income effect*).

sunrise industry See *infant industry*.

supply Indicates the various quantities of a good that firms (or a firm) are willing and able to produce and sell at different possible prices during a particular time period, *ceteris paribus* (all other things being equal).

supply curve A curve showing the relationship between the price of a good and the quantity of the good supplied, *ceteris paribus* (all other things equal); see *supply*.

supply of money The amount of money in circulation, determined by the central bank of a country; in combination with the demand for money, the supply of money determines the equilibrium rate of interest. (In practice central banks have difficulties in accurately controlling the supply of money.)

supply-side policies A variety of policies that focus on aggregate supply, namely factors aiming to shift the long-run aggregate supply ($LRAS$) curve to the right, in order to achieve long-term economic growth. They do not attempt to stabilise the economy (i.e. to reduce the severity of the business cycle). There are two major categories of supply-side policies: market-based and interventionist. To be contrasted with *demand-side policies*.

surplus In general, this is the excess of something over something else to which it is being compared. (i) In the context of demand and supply, it is the extra supply that results when quantity supplied is greater than quantity demanded. (ii) In the case of consumer and producer surplus, it is the extra benefit consumers get by paying less for a good than the amount they are willing to pay, or the extra benefit producers get by receiving a higher price for the good they are selling than the price they are willing to receive. (iii) In the case of the government budget, a surplus occurs when government revenues are greater than government expenditures. (iv) In the balance of payments, a surplus in an account occurs when the credits (inflows of money from abroad) are larger than the debits (outflows of money to other countries).

sustainability One of the key concepts of this course; refers to maintaining the ability of the environment and the economy to continue to produce and satisfy needs and wants into the future for future generations; depends crucially on the preservation of the environment over time. Related to the concept of sustainable development, meaning 'Development which meets the needs of the present without compromising the ability of future generations to meet their own needs' (according to the Brundtland Commission). *One of the key concepts of this course*

sustainable debt A level of debt where the borrowing government has enough revenues to meets it debt obligations (payment of interest and repayment of borrowed amount) without overdue debt payments, while also allowing economic growth at an acceptable rate.

sustainable development Development involving the use of resource in the present to meet present needs and wants in ways that do not deplete or degrade them, so that future generations will have enough resources to meet their own needs; refers to growth and development that does not deplete or degrade resources.

Sustainable Development Goals (SDGs) A set of 17 goals developed at the United Nations Conference on Sustainable Development in Rio de Janeiro in 2012; continue and expand upon the work begun years earlier by the Millennium Development Goals (MDGs), which ran until 2015. Are used by international organisations and national governments in their fight against poverty and efforts to achieve sustainable economic development.

tariffs Taxes on imported goods; they are the most common form of trade restriction. Tariffs may serve two purposes: to protect a domestic industry from foreign competition (a protective tariff); or to raise revenue for the government (a revenue tariff).

theory A general explanation of a set of interrelated events, usually (though not always) based on several hypotheses that have been tested successfully tire to explain why certain events happen.

theory of absolute advantage If countries specialise in and export the goods in which they have an *absolute advantage* (can produce with fewer resources), there results an improvement in resource allocation and increased production and consumption in each country.

theory of comparative advantage As long as opportunity costs in two (or more) countries differ, it is possible for all countries to gain from specialisation and trade according to their *comparative advantage*; this results in an improvement in the global allocation of resources, resulting in greater global output and consumption. Is a more powerful explanation of the gains from trade than the theory of absolute advantage.

total costs The total costs incurred by a firm that undertakes production of something.

total product The total quantity of output produced by a firm.

total revenue The amount of money received by firms when they sell a good (or service); it is equal to the price (P) of the good times the quantity (Q) of the good sold. Therefore total revenue = $P \times Q$.

tradable permits Permits that can be issued to firms by a government authority or an international body, and that can be traded (bought and sold) in a market, the objective being to limit the total amount of pollutants emitted by the firms. If a firm can produce its product by emitting a lower level of pollutants than the level set by its permits, it can sell its extra permits in the market. If a firm needs to emit more pollutants than the level set by its permits, it can buy more permits in the market.

trade creation The replacement of higher cost products (imported or domestically produced) by lower cost imports that results when a trading bloc is formed and trade barriers are removed. (To be contrasted with *trade diversion*.)

trade diversion The replacement of lower cost products (imported or domestically produced) by higher cost imports that results when a trading bloc is formed and trade barriers are removed. (To be contrasted with *trade creation*.)

trade liberalisation The policy of liberalising (freeing up) international trade by eliminating trade protection and barriers to trade (i.e. tariffs, quotas, etc.).

trade protection Government intervention in international trade through the imposition of trade restrictions (or barriers) to prevent the free entry of imports into a country and protect the domestic economy from foreign competition.

trading bloc A group of countries that have agreed to reduce tariff and other barriers to trade for the purpose of encouraging the development of free or freer trade and co-operation between them. See also free trade area, customs union and common market.

tragedy of the commons A story about cattle grazing on commonly owned land, illustrating *rivalry* and *non-excludability* of common pool resources as the land becomes overused and degraded.

transactions in non-produced, non-financial assets Refers to one of the two items in the *capital account* of the *balance of payments*, including the purchase or use of natural resources that have not been produced (land, mineral rights, forestry rights, water, fishing rights, airspace and electromagnetic spectrum); see *capital account*.

transfer payments Payments made by the government to individuals specifically for the purpose of redistributing income, thus transferring income from those who work and pay taxes towards those who cannot work and need assistance. Groups receiving transfer payments may include older people, sick people, very poor people, children of poor families, unemployed people and others; referred to as 'vulnerable groups'.

undervalued currency A currency whose value is lower than its free-market value; may occur if the exchange rate is fixed (or pegged), or in a managed exchange rate system, but not in a freely floating exchange rate system. To be contrasted with *overvalued currency*.

unemployment rate A measure of the amount of unemployment in an economy, expressed as a percentage, calculated by taking the total number of unemployed people in an economy and dividing by the labour force, and multiplying by 100.

unemployment The number of unemployed people, defined as all people above a particular age (i.e. not children) who are not working and who are actively looking for a job.

unfair competition Practices that countries may use in order to gain a competitive advantage over other countries in order to unfairly increase their exports at the expense of other countries, for example maintaining an undervalued currency.

unitary *PED* A price elasticity of demand value of one; see *price elasticity of demand*.

unitary *PES* A price elasticity of supply value of one; see *price elasticity of supply*.

universal basic income A method intended to provide everyone in a country with a sum of money that they would receive regardless of any other income they may have; its purpose is to reduce income inequalities and poverty.

unsustainable production Production that uses resources unsustainably, leading to their depletion or degradation; see *sustainability.*

utility A subjective concept, it is the satisfaction that consumers gain from consuming something.

wage A payment, per unit of time, to those who provide labour; this includes all wages and salaries, as well as supplements (such as bonuses and commissions).

wealth The money or things of value that people own minus debts to banks or other financial institutions. May include savings deposits (money saved in a bank); stocks in the stock market; bonds; land, houses and other property; valuable paintings or jewellery.

wealth taxes Taxes on ownership of wealth, most commonly *property taxes*, based on the value of property owned, and inheritance taxes, based on the value of property inherited.

weighted price index A measure of average prices in one period relative to average prices in a reference period called a base period; a weighted price index is a price index that 'weights' the various goods and services according to their relative importance. In the consumer price index (CPI), goods and services are weighted according to their relative importance in consumer spending.

welfare In general, refers to the well-being of a population. In microeconomics, it is measured by the amount of social surplus (consumer and producer surplus) that is generated in a market. Welfare is greatest, i.e. social surplus is greatest, in competitive market equilibrium when there are no externalities, and marginal social benefits are equal to marginal social costs ($MSB = MSC$).

welfare loss Refers to loss of a portion of social surplus that arises when marginal social benefits are not equal to marginal social costs ($MSB \neq MSC$), due to market failure.

what/how much to produce One of the three basic economic questions; refers to the choice that must be made in response to the question what particular goods and services an economy is to produced and in what quantities.

World Bank A development assistance organisation, composed of 189 member countries which are its joint owners, that extends long-term credit (loans) to developing country governments for the purpose of promoting economic development and structural change.

World Trade Organization (WTO) An international organisation that provides the institutional and legal framework for the trading system that exists between member nations worldwide, responsible for liberalising trade, operating a system of trade rules and providing a forum for trade negotiations between governments, and for settling trade disputes.

Index

Note: Page numbers in **bold** indicate definitions of key–terms

C

D

E

F

G

H

I

M

N

Q

R

S