1

Microeconomics

1

1

Individual economic decision making

This chapter develops from our explanation of demand theory in Book 1, section 2.1, 'The determinants of demand for goods and services', on pages 21–30. We begin with a short recap of the parts of demand theory that you must know in order to understand this introductory chapter of Book 2. Following the recap, the chapter then introduces you to two extremely significant aspects of individual economic decision making which are not covered in Book 1. The first is utility theory, an old-established body of theory which underlies the development of demand theory. Then, following a brief discussion of how imperfect information affects individual decision making, we introduce you to the second important part of this chapter, behavioural economics. Behavioural economics is a relatively new part of the subject which provides significant insights into how individuals make economic decisions.

LEARNING OBJECTIVES
This chapter will:

- remind you of some of the main elements of demand theory introduced in Book 1, Chapter 2
- discuss the significance of utility maximisation for individual economic decision making
- explain the importance of the margin when making choices
- discuss how imperfect information and asymmetric information affect choice decisions
- outline the emergence of behavioural economics as an important recent development in economic theory
- investigate important elements of behavioural economics such as bounded rationality, biases in individual decision making and the role of altruism
- relate behavioural economics to government economic policy
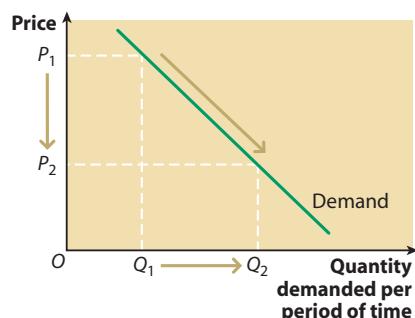
1.1 Consumer behaviour

Demand theory revisited

In this and the following paragraphs we are not going to repeat the whole of the demand theory we explained in Book 1. Rather, we shall focus solely on the elements of demand theory which are relevant to individual economic decision making.

With this in mind, we focus on **individual demand** rather than market demand. But to remind you, a market demand curve shows how much of

Figure 1.1 An individual's demand curve

Figure 1.2 A rightward shift of demand

a good or service *all* the consumers in the market plan to demand at all the different possible prices of the good or service, whereas an individual's demand curve shows how much a single consumer in the market plans to demand at all the different possible prices of the good or service. (Remember that the market demand curve is simply the sum of all the individual demand curves in the market.)

We introduced you in Book 1 to the **law of demand**, which states that as a good's price falls, more is demanded. An individual's demand curve thus shows an inverse relationship between price and quantity demanded. This relationship is shown in Figure 1.1, which is a repeat of Book 1, Figure 2.1, except that in this case the demand curve shows how a single consumer behaves in the market rather than all consumers taken together.

Having explained the 'law' of demand in Book 1, we then went on to distinguish between a **shift of a demand curve** to a new position (an **increase in demand** or a **decrease in demand**) and a movement or adjustment along a demand curve, in response to a change in the good's price. The latter we called an **extension of demand** or a **contraction of demand**. A demand curve will shift if any of the factors influencing demand, *other than the good's own price*, changes. These factors, which are sometimes called the **conditions of demand**, include income, tastes and preferences, and the prices of substitute goods and complementary goods. An increase in income shifts demand curves rightward — but only for normal goods. A normal good is defined as a good for which demand increases when income increases. By contrast, an inferior good is a good (such as poor-quality food) for which demand falls as income increases. If the good is inferior, an increase in income shifts the demand curve leftward. Figure 1.2 below shows a rightward shift of demand from D_1 to D_2, caused perhaps by a fall in the price of a good in joint demand (a complementary good) or by a successful advertising campaign for the product.

3

Rational economic decision making and economic incentives

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

Given the assumption of rational economic behaviour, a change in the price of any good and a change in the conditions of demand (and/or supply), which leads to a change in price, alters the economic incentives facing a consumer. As we have seen, with a traditional downward-sloping demand curve, a fall in the price of a good, relative to the prices of other goods, creates the incentive to demand more of the good. Likewise, an increase in the good's relative price creates an incentive to demand less of the good.

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

> **SYNOPTIC LINK**
>
> At this point, go back to Book 1, pages 25–30, and remind yourself, first, of how on occasion an individual's demand curve may slope upward, and second, of how price elasticity of demand and cross elasticity of demand affect the incentives consumers face when prices change.

> **TEST YOURSELF 1.1**
>
> Which of the following provides the best reason why consumers become early adopters of a new innovation such as a smart watch, even though they know that they will be paying a high price for the good?
>
> Early adopters are people who:
>
> **A** like technological gadgets
>
> **B** get up early to buy in a sale
>
> **C** base their consumption decisions on the reviews submitted online by existing users
>
> **D** want to be the first to get new types of product as they come onto the market
>
> Explain your answer.

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

What is utility?

We mentioned in the previous section on rational economic decision making that consumers attempt to maximise the welfare or utility they gain from the goods and services they decide to consume. We shall explore this further in the next section, on utility maximisation. In economics, **utility** is usually defined as the pleasure or satisfaction obtained from consumption.

> **KEY TERMS**
>
> **utility** the satisfaction or economic welfare an individual gains from consuming a good or service.
>
> **marginal utility** the additional welfare, satisfaction or pleasure gained from consuming one extra unit of a good.

○ **TEST YOURSELF 1.2**
A family's typical weekly shopping basket might include 'pleasure items' such as packets of crisps and cans of Coca-Cola and other items which fulfil a need and without which life would be more uncomfortable. Into which category would you place:
● medicine
● chocolate
● daffodil bulbs
● electric light bulbs
● washing-up liquid?

The relationship between total utility and marginal utility

Let us imagine a thirsty child who drinks six glasses of lemonade on a hot sunny afternoon, deriving successively 8, 6, 4, 2, 0 and –2 'units of utility' from each glass consumed. This information is shown in the total and marginal utility schedules in Table 1.1, from which the total and marginal utility curves drawn in Figure 1.3 are plotted.
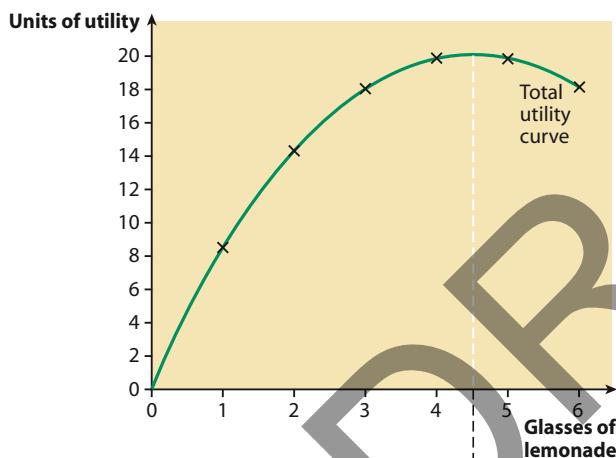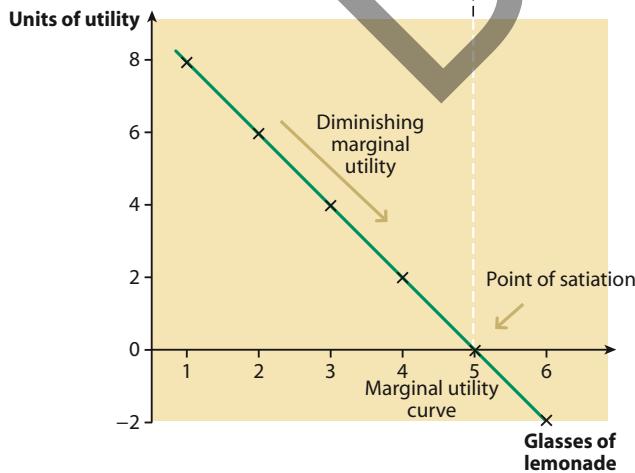

Figure 1.3 An example of total utility and marginal utility curves

Table 1.1 Total and marginal utility schedules for lemonade

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

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

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

5

Economists refer to the utility of a good: in this case, how much satisfaction can be received from consuming glasses of lemonade

In Figure 1.3, diminishing marginal utility is shown both by the diminishing rate of increase of the slope of the total utility curve drawn in the upper panel of the diagram and by the negative or downward slope of the marginal utility curve in the lower panel. Notice that we have drawn a 'point of satiation' on the diagram, which is reached as the fifth glass of lemonade is drunk. The fifth glass of lemonade yields zero marginal utility. At this point, when marginal utility is zero, total utility is maximised. In the context of food and drink, satiation means being 'full up'. Even if lemonade is free to the consumer, it would be irrational for our 'no-longer-thirsty' child to drink a sixth glass of lemonade. He or she would experience negative marginal utility (or marginal disutility), which is shown by the downward slope of the total utility curve and by the negative position of the lower section of the marginal utility curve.

QUANTITATIVE SKILLS 1.1

Worked example: calculating marginal utility

An 8-year-old boy decides to enter a competition to see how many jam doughnuts can be eaten in 15 minutes. Table 1.2 shows how many he ate and his total utility schedule.

Table 1.2 Total utility for doughnuts

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

Complete the boy's marginal utility schedule.

The boy's marginal utility schedule is shown in Table 1.3.

Table 1.3 Total and marginal utility for doughnuts

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

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

The numerical examples in Tables 1.1–1.3, and the graph in Figure 1.3, illustrate a very famous economic hypothesis, which some would call an economic law: the **hypothesis of diminishing marginal utility**. This simply states that as a person increases consumption of a good — while keeping consumption of other products constant — there is a decline in the marginal utility derived from consuming each additional unit of the good.

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

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

CASE STUDY 1.1

Adam Smith's diamonds and water paradox

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

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

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

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

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

Follow-up questions

1 Define the terms 'scarcity' and 'marginal utility'.
2 Can you think of two other goods which generally illustrate the paradox of value?

EXTENSION MATERIAL

Marginal utility and an individual's demand curve

If lemonade were available completely free (at zero price), it would be rational for our thirsty boy to drink exactly five glasses of lemonade in the course of a hot, sunny afternoon. He would consumer up to the point of satiation, beyond which no further utility can be gained. But because lemonade is an economic good which is scarce in supply and which has an opportunity cost, it is reasonable to assume that the child (or his parents) must pay for his drinks. Suppose that the price of lemonade is equal to the marginal utility gained from the fourth glass. At this price, P_4 represents the opportunity cost of the fourth glass of lemonade: that is, the utility that could be gained if the price were spent on some other good, say a bar of chocolate. To maximise utility at this price, the thirsty child should drink four glasses of lemonade, but no more. It would be irrational to consume a fifth glass at this price, since the extra utility gained would be less than the opportunity cost represented by the price P_4.

Figure 1.4 below shows the effect of the price rising from P_4, successively to P_3, P_2 and P_1. These prices equal the marginal utility derived by the child from the third, second and first glasses of lemonade. When the price rises to P_3, our thirsty child reduces demand to three glasses, so as to maximise utility in the new situation. At price P_2 demand is again reduced to two drinks, and so on. The higher the price, the lower the quantity demanded, which is exactly what a demand curve shows.
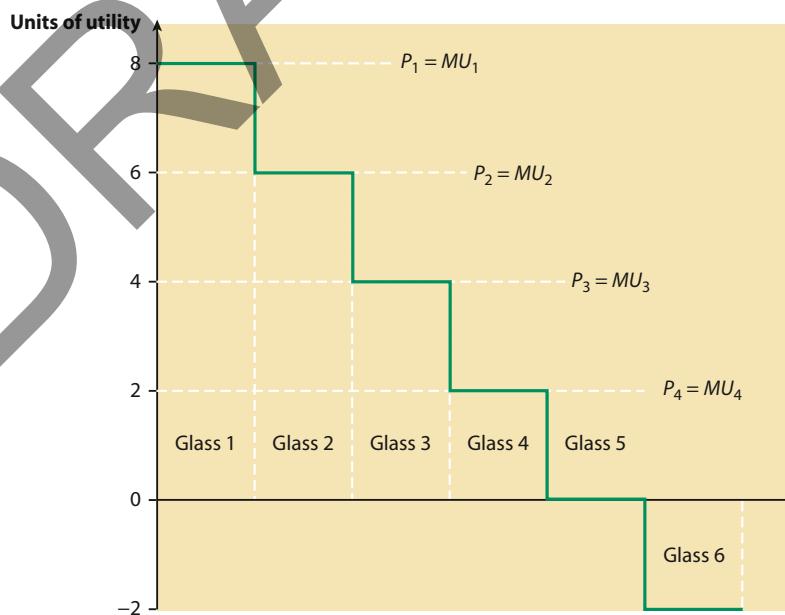
Figure 1.4 Relating marginal utility to changes in price and to the shape of a demand curve

Utility maximisation

As we noted in Book 1, the assumption of maximising behaviour by economic agents (consumers, workers, firms and even the government) is central to orthodox or traditional economic theory. Economic agents decide their market plans so as to maximise a target objective or goal which is believed to be consistent with the pursuit of self-interest. In demand theory, the objective which households are assumed to wish to maximise is the utility, or satisfaction, obtained from the set of goods and services consumed.

EXTENSION MATERIAL

Maximising versus minimising behaviour

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

Maximisation subject to constraints

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

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

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

Importance of margin when making choices

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

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

EXTENSION MATERIAL
Can utility be measured?

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

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

CASE STUDY 1.2

The November 2012 edition of the *Economic Review*, a magazine published by Philip Allan for Hodder Education, included an article 'The economics of happiness' written by Corrado Giulietti and Juan David Robalino. This case study is an extract from their article. If your school subscribes to the *Economic Review*, you can gain access over the internet to the complete article.

The economics of happiness

Happiness has been the subject of study of philosophy, psychology, anthropology and many other disciplines for a long time, but only recently have economists become interested in happiness.

The principal objective of public policy is to use limited resources to maximise the welfare of individuals. But how do we measure welfare or wellbeing? Economists have developed the concept of the utility function, which is a way to quantify individuals' welfare using mathematical tools. However, a major practical drawback is that utility cannot be observed directly, so how do we know whether one individual is better off than another?

For many years, when evaluating policy interventions, governments relied on *objective* measures such as the income or employment status of an individual. These indicators can capture a significant share of individuals' utility and so it was thought that utility could be maximised through maximising these objective measures.

However, our welfare comprises of many aspects that cannot be captured by economic indicators, such as social and family relationships or pollution. Studies on the economics of happiness argue that the happiness experienced by individuals can be used to approximate utility. The most common approach is to ask people how happy they are by means of a detailed set of questions. Self-reported measures are referred to as *subjective* wellbeing measures.

Subjective wellbeing is usually measured using indices derived from surveys such as the British Household Panel Survey (BHPS). Respondents are asked questions about their life satisfaction, mental health and experienced happiness, from which it is possible to derive alternative measures of wellbeing. The General Health Questionnaire (GHQ) of the BHPS consists of 12 questions about mental health, ranging from the individual's ability to 'concentrate' and 'face problems' to whether they feel 'overall happy' with their life. However, some economists are sceptical about the subjective

nature of happiness. Since they are self-reported measures of wellbeing, how comparable are they across individuals?

We might be inclined to think that happiness decreases as an individual gets older. Strikingly, this is not the case; in fact there is a U-shaped relationship between happiness and age. The GHQ index reaches its lowest score when individuals are about 45 years old. One possibility is that individuals in their mid-40s are particularly worried about achieving their career goals. More generally, this could be seen as the presence of a 'mid-life crisis'. On the other hand, after the age of 45, individuals' wellbeing increases. One potential explanation is that as they get older, individuals accept their life as it is and start enjoying it again.

Physical health produces a more straightforward result. As we might expect, people with excellent self-reported health levels are happier than those with poor health. In many studies, the relationship between health and happiness is stronger than with any other determinants of happiness.

Other major findings concern gender and marital status. Females report higher happiness than males (at least in the UK), married people are usually happier than those who are single, while divorced people are usually the least happy.

Surveys of subjective wellbeing have revealed that married people are usually happier than those who are single and that divorced people are often the least happy

If we compare employed and unemployed individuals with the same income (say, because the unemployed receive benefits from the government), the latter are less happy. This suggests that a job itself is associated with higher wellbeing. By working, individuals feel socially included and they don't perceive the social stigma of being jobless. People never adjust to unemployment. It doesn't matter how frequently or for what period of time a person is jobless: unemployed individuals remain persistently unhappy.

A puzzling result emerges from the relationship between income and happiness. Studies have found a positive relationship between happiness and income, but surprisingly this association is not very strong. The level of happiness of individuals who earn over £2500 per month is less than one point higher than those who earn less than £1000. This is named the *Easterlin paradox* after economist Richard Easterlin. So how do we explain the fact

that happiness has not increased very much over time despite material wellbeing improving for many people? One explanation is that it is not just our income that makes us happy but also how our income compares with that of people similar to us (our neighbours or classmates, say). In other words, it seems that people are relatively happy when they earn more than the people around them. Therefore, if everybody gets richer by the same amount, there is no substantial increase in the level of happiness.

Follow-up questions

1 Do you think that governments should rely on subjective measures of wellbeing rather than objective indicators such as the level of real GDP to evaluate the effectiveness of public policy? Justify your answer.
2 Research what the current UK government has been saying about the economics of happiness.

1.2 Imperfect information

The importance of information for decision making

So far in this chapter, we have assumed that consumers possess perfect information — for example, about the goods that are available to buy, their prices and quality, and about the utility which will be derived from their consumption.

However, when attempting to maximise total utility, more often than not consumers possess imperfect information. As a result, they make 'wrong' decisions. We saw in Book 1, Chapter 5, section 5.5 how consumers may choose to under-consume a merit good such as education and over-consume a demerit good such as tobacco because they possess imperfect information about the long-term consequences of their choices. We shall investigate this further in Chapter 6, section 6.1 of this book, and we shall also touch on this issue in our coverage of behavioural economics in the next section of this chapter.

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

SYNOPTIC LINK
In Chapter 6 we extend the analysis of merit goods and demerit goods begun in Book 1, Chapter 5 by examining the information problems households experience when deciding how much to consume of a merit good such as education, or a demerit good such as tobacco.

ACTIVITY
Get together with a group of your fellow students and discuss the things which make you happy and the ways in which your behaviour is affected by imperfect information as you go about your daily activities. Summarise the group's results.

The significance of asymmetric information

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

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

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

CASE STUDY 1.3

The market for lemons

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

On receiving his Nobel Prize, Akerlof said:

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

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

From time to time one hears either mention of or surprise at the large price difference between new cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a 'new' car.

We offer a different explanation. Suppose that there are just four kinds of cars. There are new cars and used cars. There are good cars and bad cars (which in America are known as 'lemons'). A new car may be a good car or a lemon, and of course the same is true of used cars.

The individuals in this market buy a new automobile without knowing whether the car they buy will be good or a lemon. After owning a specific car, however, for a length of time, the car owner can form a good idea of the quality of this machine.

An asymmetry of information has developed: for the sellers have more knowledge about the quality of a car than the buyers. But good and bad used cars must still sell at the same price, since it is impossible for a buyer to tell the difference between a good and a bad car.

It is apparent that a used car cannot have the same valuation as a new car — if it did, it would clearly be advantageous to trade a lemon at the price of a new car, at a high probability of the new car being a good car. Most used cars traded will be 'lemons', and good used cars may not be traded at all. The 'bad cars tend to drive out the good (in much the same way that bad money drives out the good).'

Follow-up questions

1 Akerlof asked the question 'if he wants to sell that horse, do I really want to buy it?'. Explain how this question relates to the market for second-hand cars.
2 Suggest two other markets, other than the markets for second-hand cars, used computers and horses, in which market outcomes are affected by asymmetric information.

1.3 Aspects of behavioural economic theory

Emergence of behavioural economics

Behavioural economics is a field of study that has attracted a great deal of attention since the beginning of the 21st century. Most of the research in the field has come from universities in the USA, but in recent years UK university economics departments have been offering courses in the subject.

Behavioural economics is built on the insights of psychologists seeking to understand human behaviour and decision making. This research field can be traced back to 1931 when L. L. Thurstone conducted experiments to determine consumer preferences by asking participants to choose repeatedly between alternative bundles of goods. Two of the most influential psychologists are the Israeli academics Amos Tversky and Daniel Kahneman, who spent decades studying how people think and provided a major contribution to decision research and psychological theory. In 2002 Kahneman was awarded the Nobel Memorial Prize in economics and in 2011 he published *Thinking, Fast and Slow* in which he credits Tversky (who died in 1996) for helping with much of his work.

In 2008 the Chicago economist Richard Thaler and the legal scholar Cass Sunstein published *Nudge: Improving Decisions about Health, Wealth and Happiness*, which is a highly accessible overview of behavioural economics. Having read *Nudge*, immediately on becoming prime minister in 2010, David Cameron set up the UK government's Behavioural Insights Team (BIT), which was initially based in the Cabinet Office in Downing Street. The creation of the BIT, and of a similar body advising the US president, marked the growing influence that behavioural economics was having on government policy-makers.

On its website, the BIT writes:

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

The BIT's website is worth reading to find out about projects that the team run in government. The web address is: **http://www.behaviouralinsights.co.uk**. You might also access a BIT publication, *Better Choices: Better Deals*, otherwise known as the Government's consumer empowerment strategy, which recommends how government policy can attempt to influence consumer behaviour. It can be found at **https://www.gov.uk/government/publications/better-choices-better-deals-behavioural-insights-team-paper**. The case study below has been adapted from the policy document.

CASE STUDY 1.4

Better Choices: Better Deals

We are trying to shift power to citizens and communities. Three changes that are helping to make this possible are:

- The increasing role of new technologies, in particular internet and mobile phone applications, that have opened up new channels for consumers to find, compare, and purchase goods and services.
- The use of data, drawn from customers' own transaction histories, that have allowed businesses to understand their customers better, allowing them to make more tailored recommendations.
- The development of new ways for different consumers to collaborate across the economy — for example whether by sharing cars or bicycles, or giving feedback about a GP practice, a local tradesman or a multinational corporation.

Putting power in the hands of consumers so that they are better able to choose between suppliers, will both enable them to get the best deals for themselves individually and collectively, while also putting pressure on businesses to be more efficient and innovative. We see two profound changes taking place:

- A shift away from a world in which certain businesses tightly control the information they hold about consumers, towards one in which individuals, acting alone or in groups, can use their data or feedback for their own or mutual benefit.
- A shift toward an environment in which individuals and groups feel more able to send the right signals to business, and hence secure the products and services they want.

In short, we want to see confident, empowered consumers able to make the right choices for themselves — to get the best deals, demand better products or services, and be able to resolve problems when things go wrong. This approach makes it easier for honest, high-quality businesses to compete and will drive innovation, competition and growth. A better deal for consumers and the economy means a better deal all round.

We are setting up six new programmes to give consumers richer, more relevant information about the goods and services that they use. We are:

- Introducing Annual Credit Card Statements, containing information about fees and how to switch.
- Working with energy suppliers to provide clearer information about the lowest available energy tariff.
- Reforming Energy Performance Certificates so they include clear information about the costs of heating a home.
- Working with the Food Standards Agency, trialling new ways to help consumers understand food hygiene ratings of restaurants.
- Reforming car labelling by supporting work to give consumers clearer information about the costs of running different cars.
- Facilitating the launch of a new programme of work to examine product information in relation to health and the environment.

We will make use of the 'Power of the Crowd' by introducing a range of new initiatives that will support the development of collective purchasing and collaborative consumption. We will be:

- encouraging collective purchasing deals.
- piloting a green collective purchasing scheme in which B&Q will help encourage the uptake of energy efficiency measures.

Follow-up questions

1 The passage states 'new technologies, in particular internet and mobile phone applications…have opened up new channels for consumers to find, compare, and purchase goods and services'. Give two examples of this happening for yourself, your friends or your family.

2 An important element of *Better Choices: Better Deals* is 'the power of the crowd'. This means that consumers acting collectively can counter to some extent the market power of the businesses which sell them goods and services. Is there any evidence of this happening since *Better Choices: Better Deals* was published in 2011?

The use of the internet affects consumer behaviour in which of the following ways?

A It facilitates the use of price comparison websites

B It stops consumers comparing prices in store

C It generally adds to consumer confusion

D It slows down consumer decision making

Explain your answer.

EXTENSION MATERIAL

Squaring the circle between traditional and behavioural economic theory

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

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

However, in a very famous essay, *The Methodology of Positive Economics*, published in 1953, the great pro-free market economist Milton Friedman defended the traditional approach. Friedman wrote: 'Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions.'

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

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

CASE STUDY 1.5

In the April 2010 edition of the *Economic Review*, published by Philip Allan for Hodder Education, David Gill presented an overview of interesting developments in behavioural economics. This case study summarises the introduction of David Gill's article. If your school subscribes to the *Economic Review*, you can gain access over the internet to the complete article.

Beyond homo economicus

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

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

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

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

Third, humans clearly face quite substantial limitations of computation and reasoning. These are particularly important when the environment is complex — e.g. when people are interacting strategically, my best action will depend on what others choose to do. An oligopolistic firm deciding what price to set will have to start thinking about what their competitors are going to do.

Follow-up questions

1 What is meant by altruism, and how does altruism affect the economics of charities?
2 What is a 'rule-of-thumb'? Give an example of a rule-of-thumb you often use.

KEY TERMS

bounded rationality when making decisions, an individual's rationality is limited by the information they have, the imitations of their minds, and the finite amount of time available in which to make decisions.

bounded self-control limited self-control in which individuals lack the self-control to act in what they see as their self-interest.

Rational economic behaviour revisited

Bounded rationality

So far in this chapter, and also in Book 1, we have assumed that when exercising choice, individuals are perfectly rational, in the sense that they make decisions in a context of being fully informed, with perfect logic and aiming to achieve the maximum possible economic gain. However, in real life, individuals are seldom if ever perfectly rational. In the world in which we live, decisions are made in conditions of **bounded rationality**, which means that individuals, however high or low their intelligence, make decisions subject to three unavoidable constraints: imperfect information about possible alternatives and their consequences; limited mental processing ability; and a time constraint which limits the time available for making decisions. In complex choice situations, bounded rationality often results in *satisficing* rather than *maximising* choices.

> ## SYNOPTIC LINK
> The difference between satisficing and maximising is explained in Chapter 3, pages XX–XX.

Bounded self-control

Bounded rationality is closely linked to the related concept of **bounded self-control**. Traditional or orthodox economic theory implicitly assumes that when making choices, individuals have complete self-control. Behavioural economists, by contrast, believe that individuals have bounded (or limited) self-control. Making New Year resolutions in the period immediately after Christmas provides many good examples. Having put on weight during the Christmas festivities, people may decide to go for a daily jog early in the morning before going to work each day after 1 January. For many, this may work well for a few days, but the first bout of bad weather often leads to the resolution being broken.

Thinking fast and thinking slow

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

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

For example, when learning to play a new game such as golf, an individual will 'think slow' when deciding on the appropriate golf club to select for a particular stroke, and on how to grip the club and to take a swing at the ball. Because the decision making is relatively slow, involving careful, logical thought about every decision, the process can be tiring. However, the more often the game is played and the more practice is put in, the less

will golfers have to think about minor decisions. Automatic thinking takes over. Professional golf players often play quickly and instinctively. Through years of repetitive training, their automatic systems have learnt to respond to situations promptly and effectively. In big-game situations they can, of course, suffer if they stop to think. When this happens, they revert back to System 1 or the reflective system, which can mean that bad decisions lead to disastrous consequences.

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

ACTIVITY

Make a list of all the things you bought the last time you went on a serious shopping expedition. How many of your decisions to buy were undertaken by your reflective system and how many by your automatic system?

CASE STUDY 1.6

Daniel Kahneman

Daniel Kahneman pretty much created the field of behavioural economics. His central message could not be more important, namely, that human reason left to its own devices is apt to engage in a number of fallacies and systematic errors, so if we want to make better decisions in our personal lives and as a society, we ought to be aware of these biases and seek workarounds. That's a powerful and important discovery.

Steven Pinker, psychology professor at Harvard University

Along with Amos Tversky, Daniel Kahneman is famous for researching the apparently strange way in which people make decisions in risky situations. Kahneman and Tversky realised that people behaved in different ways depending on how the risky situation was presented. If a risk is presented in terms of losses, people will be more risk seeking, but if it is expressed in terms of gains, people will be more risk averse.

Their classic example involves this fictional situation:

Imagine your country is preparing for the outbreak of a disease expected to kill 600 people. If program A is

adopted, exactly 200 people will be saved. If program B is adopted there is a 1/3 probability that 600 people will be saved and a 2/3 probability that no people will be saved.

Here, the risk is presented in terms of gains, and 72% of people tend to choose option A. Here's the same problem but this time presented in terms of losses:

Imagine your country is preparing for the outbreak of a disease expected to kill 600 people. If program A is adopted, exactly 400 people will die. If program B is adopted there is a 1/3 probability that no one will die and a 2/3 probability that 600 people will die.

Now 78% of people choose B because the problem is presented in terms of losses. People suddenly prefer to take a risk. The two situations are in fact mathematically identical, yet people's decisions are heavily influenced by the way the problem is *framed*. This effect has been termed *preference reversal*.

Follow-up questions

1 Explain the difference between 'risk seeking' and 'risk aversion'.
2 The final paragraph states that 'people's decisions are heavily influenced by the way the problem is *framed*'. Explain what this means.

19

Biases in decision making

KEY TERM

cognitive bias a mistake in reasoning or in some other mental thought process occurring as a result of, for example, using rules-of-thumb or holding onto one's preferences and beliefs, regardless of contrary information.

Behavioural economics argues that the decisions people make when exercising choice are often heavily biased. This is because decisions are made on the basis of one's own likes, dislikes and past experiences. Psychologists use the term **cognitive bias** to describe this situation.

A cognitive bias is a mental error that is consistent and predictable. There are many kinds of cognitive bias, one of which is *confirmation bias*. This is the tendency to seek only information that matches what one already believes. It stems from the often unconscious act of listening only to opinions which back up our pre-existing views, while at the same time ignoring or dismissing opinions — no matter how valid — that threaten our views.

EXTENSION MATERIAL

The AQA specification advises that you understand some of the reasons why an individual's economic decisions may be biased. Here are ten examples of cognitive bias additional to confirmation bias, which we have already described:

Ten cognitive biases

Status-quo bias: This is where people generally prefer that things remain the same, or change as little as possible. It is the belief that changing the status quo is likely to be inferior or make things worse.

Memory bias: People are likely to possess accurate memories associated with significant emotions or events (such as the memory of what one was doing when a grandchild was born or when a catastrophe such as the assassination of a world leader occurred). Memory bias influences what and how easily one remembers.

Observational selection bias: The effect of suddenly noticing things not noticed much before and wrongly assuming that the frequency of the observation has increased.

In-group bias: We forge tighter bonds with friends similar to ourselves (our in-group), while being suspicious of others. We value our immediate group at the expense of people we don't really know.

Positive expectation bias: The sense that luck will eventually change for the better — which often fuels gambling addictions. A run of bad luck has to change eventually and better times lie ahead.

Post-purchase rationalisation: Believing after the purchase of an unnecessary, faulty or overly expensive good that buying the product was a good idea, thus justifying a bad decision.

Neglecting probability: The inability to grasp a proper sense of peril and risk, which can lead to overstating the risks of relatively harmless activities and understating the risks of more dangerous ones — for example, air travel versus car travel or cycling risks.

Negativity bias: People tend to pay more attention to bad news than to good news. We perceive negative news as being more important or profound.

Bandwagon effect bias: People succumb to 'group-think' or herd behaviour.

Current moment bias: Preferring pleasure or gratification at the current moment to pleasure in the future. The pain can be left for later.

ACTIVITY

Make a list of all the significant decisions you have made in the last week. To what extent, if any, do these decisions embody one or more of the ten cognitive biases listed alongside?

The availability bias

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

Consider also the economic decision to buy a lottery ticket. The probability of selecting the winning numbers in a draw is outrageously long at over 14 million/1. It is irrational to believe that buying a ticket is a sound economic decision because the chance of winning the jackpot is so improbable. However, in October 2014 UK National Lottery tickets sales totalled £649.4 million, an increase of £134.0 million on the same period in 2013. No doubt, when buying a ticket most players do not think about the odds but instead focus on the news stories of people winning the jackpot. The lucky winners of large jackpots are publicised in the national media and their tales are promoted by Camelot, the business that runs the National Lottery. Since its launch in 1994 the National Lottery claims to have created 3,700 millionaires in the UK.

The availability bias often leads to decisions that are not based on logical reasoning. The media will report stories that stick in our mind and affect our reasoning process. Humans will often believe that the probability of an extreme weather event, such as a hurricane or severe flooding, is more likely than empirical statistical analysis bears out. In October 2014 Ipsos MORI's published research highlighting how the general public in 14 countries held preconceptions on the make-up of their societies that were significantly detached from the reality. In the UK, for example, the average citizen believed that 24% of the population were immigrants when the real figure is 13%; and likewise that 24% of the working age population was unemployed when in fact it was less than 7%.

Quickly recalling examples that come to mind is an automatic system response. It will often lead to an overly cautious decision that over-estimates the probability of an outcome occurring.

Anchoring

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

Biases based on social norms

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

Negative social norms include attitudes towards drinking alcohol. Many young adults often drink heavily because they think it is what people of their age are expected to do. By presenting statistical data showing that the majority of

---

KEY TERM

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

KEY TERMS

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

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

21

KEY TERMS

economic sanctions in this context, restrictions imposed by regulations and/or laws that restrict an individual's freedom to behave in certain ways. Breaking a sanction can lead to punishment.

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

young adults do not engage in regular heavy drinking, behavioural economists would seek to *nudge* young adults into different patterns of behaviour.

Positive social norms can be seen in the way in which social attitudes have altered toward smoking in the last 30 years. In the 1980s it was socially acceptable to smoke in all public places including libraries, trains and the London Underground. Concerted health campaigns which provided the general public with better information about the risks of smoking have altered social attitudes toward smoking. As a result people became much more willing to accept laws which restricted their right to smoke. The laws banning smoking in public places are **economic sanctions** (used by government policy-makers) and not **nudges**. Critics of behavioural economics point out that sanctions, such as the smoking ban, are more effective at changing behaviour and improving public health than nudges, which only alter the behaviour of some people. Nevertheless, government reports in Ireland claim that since smoking in public places was banned people are also less likely to smoke in other people's houses because it is now considered to be socially unacceptable.

SYNOPTIC LINK
The cases for and against economic sanctions and nudges are developed in Chapter 6, pages XX–XX.

ACTIVITY
Give examples of some of the social norms that affect your behaviour when at home and when attending school or college.

TEST YOURSELF 1.4
Which of the following provides the best definition of a norm?

Norms are:

A laws that attempt to discourage excessive consumption

B informal rules that govern human behaviour

C formal rules that govern human behaviour

D formal rules about how to buy goods and services

Explain your answer.

Altruism and fairness

Altruism is when we act to promote someone else's wellbeing, even though we may suffer as a consequence, either in terms of a financial or time loss, or by incurring personal risk. Before the development of behavioural economics, economists generally assumed that individuals were not altruistic and acted only in their self-interest. Nevertheless, altruism could still be accommodated within maximising theory — for example, by assuming that individuals derive pleasure as a result of giving to others. More recently, behavioural economists have drawn attention to the fact that for many if not most people, their first

KEY TERM
altruism concern for the welfare of others.

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

SYNOPTIC LINK
Go back to Book 1, Chapter 1 to remind yourself of the meaning of normative statements.

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

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

impulse is to cooperate with each other rather than to compete. Very young children are frequently observed helping other children around them, out of a genuine concern for their welfare. Animals have also been observed displaying altruism.

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

1.4 Behavioural economics and economic policy

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

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

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

Choice architecture and framing

Choice architecture

Choice architecture is the term used by behavioural economists to describe how government policy-makers can lead people into making particular choices. Government can use behavioural insights to design choice architectures so that citizens are *nudged* to opt for choices that are deemed to be in their best interest, so as to achieve a socially desirable outcome. For example, countries that require people to opt out of organ donations generally have a much higher proportion of the population willing to donate than countries that ask people to opt in.

This introduces us to a key behavioural concept known as **default choice**. When framing policy on issues such as organ donation, individuals who, in the event of their death, might donate body organs such as hearts or livers, can be asked whether to opt in or opt out of organ donation. In this context, an 'opt-in' default choice is illustrated by the use of a tick box which, if filled

in by member of the general public, indicates positively that they would like to donate their body organs after their death. Unless the user ticks the box, healthcare organisations such as the NHS cannot make use of their organs. This is in contrast with an 'opt-out', where the default position is that the organs can be used to help others survive unless the box has been ticked to indicate that the body parts should *not* be used. For a behavioural economist, the benefits of opt-out over opt-in are clear: the supply of donated organs rises to be closer to the demand for them and the nation's public health improves.
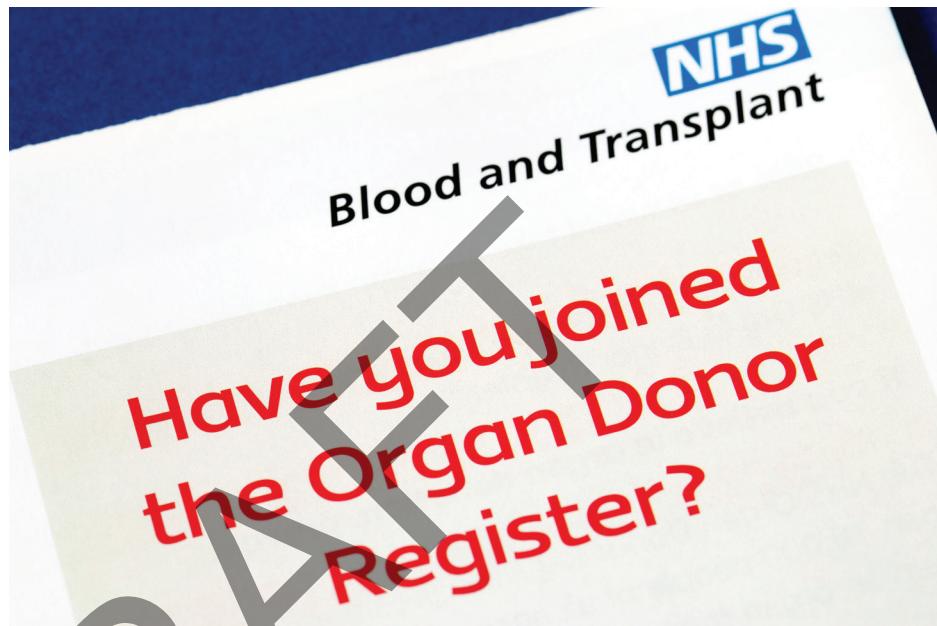
Signing up to the NHS Organ Donor Register is an opt-in default choice

Policy-makers can improve social welfare by designing government programmes that select as a default an option that can be considered in an individual's best long-term interest. A number of examples of this approach have been trailed and introduced by the UK government's Behavioural Insights Team. One of these is automatic pension enrolment (see case study 1.7).

CASE STUDY 1.7

Automatic pension enrolment

There is a broad consensus among economists in the UK that too many workers are saving too little for their retirement in old age. In an effort to solve this problem the BIT and the Department for Work and Pensions (DWP) introduced a policy of automatic enrolment in October 2012. Previously the default position for workers was that they would not pay into a pension fund unless they made the choice to opt into a scheme. Under automatic enrolment the default position is that workers pay into a pension system unless they choose to opt out.

Initially the scheme started with the UK's biggest employers (firms that have over 250 workers) but it is being rolled out to include all employers by

2018. After the first 6 months the BIT reported that overall participation in pension schemes increased from 61% to 83%. This saw 400,000 extra workers saving income for retirement. By December 2014 the DWP reported that 5 million workers were included in automatic enrolment and that 9 out of 10 workers had not exercised their right to opt out of the system.

Follow-up questions

1 Why is it desirable for social and economic policy for governments to base pension policy on an opting-out rather than an opting-in default choice?
2 Research the BIT website to find other examples of policy being changed to incorporate the opting-out default choice.

Framing

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

Politicians will often frame (or spin) economic statements in a manner that is favourable to the argument they are trying to make. For example, in December 2014 the chancellor of the exchequer George Osborne said that the government had more than halved the UK's budget deficit since taking office in May 2010. This message was printed on Conservative Party campaign posters in January 2015. Osborne was trying to frame his government in the voters' mind as one of economic competence. This statement is true if you measure the size of the budget deficit as a ratio of GDP. However, if you measure the budget deficit in money terms, it has only been reduced by around 40%.

Table 1.4 UK budget deficit, 2009/10 and 2014/15

Year	UK budget deficit (£bn)	UK budget deficit as a % of GDP
2009/10	153.0	10.2
2014/15	91.3	5.0

Source: OBR, *A Brief Guide to the UK Public Finances*, 3 December 2014

Mandated choices

A variation of default choice is **mandated choice**; this is where people are required by law to make a decision. A mandated or required choice is when a choice architect designs a system that forces individuals to make an explicit decision and not merely go ahead with a default position. This system is favoured by libertarians who philosophically oppose the notion that well-meaning government officials should guide citizens into making the 'correct' choice favoured by the government, especially if this is the default option.

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

Restricted choice

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

Government policy-makers should consider behavioural insights when designing systems. A well-designed system should make it easier for citizens to pay for government services by setting up direct debits, using accessible

language and sending text messages or e-mails to remind people to complete requests. Evidence from the BIT shows that personalised letters increase response rates, whilst asking respondents to sign forms at the top of the page and not the bottom results in more honest answers.

Choice architecture: some further implications for government policy

When making a choice, individuals need to understand the decision that they are making. Simple decisions such as ordering a meal in a restaurant are easy to understand. More complex decisions, such as taking out a mortgage or insurance policy, can be difficult to understand due to complex clauses, intricate pricing tariffs, and baffling legal terminology. Government regulation of business behaviour should try to ensure that companies make their products as straightforward and transparent as possible for consumers to exercise choice. Individuals need to have as great an understanding as possible of the consequences of any decisions and choices they make.

Well-designed choice architecture helps people to make good intuitive decisions. For example, pedestrians are told by the writing painted on the roads to *look right* or *look left* when crossing a road. This choice architecture helps reduce accidents, especially in tourist areas where a large number of pedestrians are not initially familiar with the 'rules of the road'.

However, as choices become more complex, people have greater difficulty in understanding the information presented to them. By providing an individual with information about the choices made by similar people in similar situations, it is possible for an individual to benefit from 'collaborative filtering'. Accessing the preferences of like-minded people reduces the chance that individual decisions are made on the basis of imperfect knowledge, though to some extent this advantage may be offset by people's tendency to 'join the herd' without considering the disadvantages of doing so.

The best way to help an individual's decision-making process is to provide feedback that enables them to learn from their past performance. In a school environment, good teachers do this all the time. Constructive feedback helps people to make better decisions and choices. However, negative feedback is typically misperceived or rejected. Constructive feedback enhances people's feelings of competence and self-control. Feedback is most useful when individuals actively participate in the feedback session. By contrast, destructive feedback tends to cause conflict and reduce personal motivation.

Choice architects need to build incentives — sometimes based on monetary rewards — into the choice architecture they design. People often respond to such incentives. According to traditional economic theory, individuals value money and other tangible rewards and try hard to gain them. People go to work to earn money, so we expect them to work harder when there is more money at stake. Reward incentives — particularly monetary incentives — can motivate individuals to behave in ways they would otherwise avoid.

However, behavioural economics has made two important advances with regard to reward incentives and how they affect economic behaviour. First, it has suggested that not all incentives are equally important. As we saw

earlier, individuals feel losses more severely than equivalent gains. Second, behavioural economics has shown that in certain situations, individuals respond in perverse ways to reward incentives. Monetary incentives, for example, may cause individuals to respond in the workplace with less effort rather than more. Behavioural economics recognises that people are not only motivated by financial gain; social norms and perceptions of fairness, for example, exert a powerful influence on people's behaviour.

Revisiting nudge theory

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

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

Nudges versus shoves

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

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

Table 1.5 Nudges versus shoves

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

ACTIVITY

School rules, which may not have changed significantly for many years, are often based on the 'shove' principle. These include punishments for lateness and bad behaviour. Get together with a group of fellow students and discuss how, and the extent to which, the school might move away from 'shove' to 'nudge'. Then see what happens, both immediately and in the future, when you submit your proposals to the school authorities — for example, at a School Council meeting.

CASE STUDY 1.8

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

Applying behavioural insights to charitable giving

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

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

Four behavioural insights

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

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

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

● Attracting individuals' attention, for example by using personalised messages
● Rewarding the behaviour you seek to encourage, for example through matched funding schemes
● Encouraging reciprocity with small gifts

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

● Using prominent individuals to send out strong social signals
● Drawing on peer effects, by making acts of giving more visible to others within one's social group
● Establishing group norms around which subsequent donors 'anchor' their own gifts

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

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

Follow-up questions

1 The BIT paper recommends 'establishing group norms around which subsequent donors "anchor" their own gifts'. Explain the two terms 'group norms' and 'anchor'.
2 Identify from within the extract, two examples of nudge theory being applied.

28

SUMMARY

- The starting point for understanding individual economic decision making is understanding the nature of demand, rationality and maximising behaviour.
- Economists have traditionally assumed that individuals wish to maximise utility.
- Utility can be thought of as satisfaction, pleasure or fulfilment of need.
- It is important to distinguish between total utility and marginal utility.
- The hypothesis (or 'law') of diminishing marginal utility lies behind the derivation of an individual's demand curve.
- Maximisation of utility takes place, subject to a number of constraints which include a budget constraint.
- Utility cannot be measured directly but can be indicated by revealed preference.
- Individual economic decision making is affected by imperfect and asymmetric information.
- In recent years, behavioural economics has emerged to question many of the assumptions of traditional economic theory.
- Key concepts in behavioural economics include bounded rationality, bounded self-control, biases in decision-making processes and anchoring.
- Biases are often based on social norms.
- Nudge theory, choice architecture and framing lie at the heart of the ways in which behavioural architecture can influence economic policy making.

Questions

1 Describe the main features of an individual's demand curve for a good.

2 What is meant by maximisation subject to constraints?

3 Explain the difference between maximising and satisficing behaviour.

4 Discuss the similarities of, and the differences separating, the orthodox and the behavioural theories of individual economic decision making.

5 What is a 'nudge'? Explain how a food business might use nudges to promote healthy eating.

6 Outline two ways in which the insights of behavioural economics can be incorporated into government economic policy.

2

Production, costs and revenue

To understand production, costs and revenue in greater depth than was the case in Book 1, it is necessary to understand how the 'building blocks' of the theory of the firm, which are shown in Figure 2.1, link together. This and the next chapter explain these linkages.
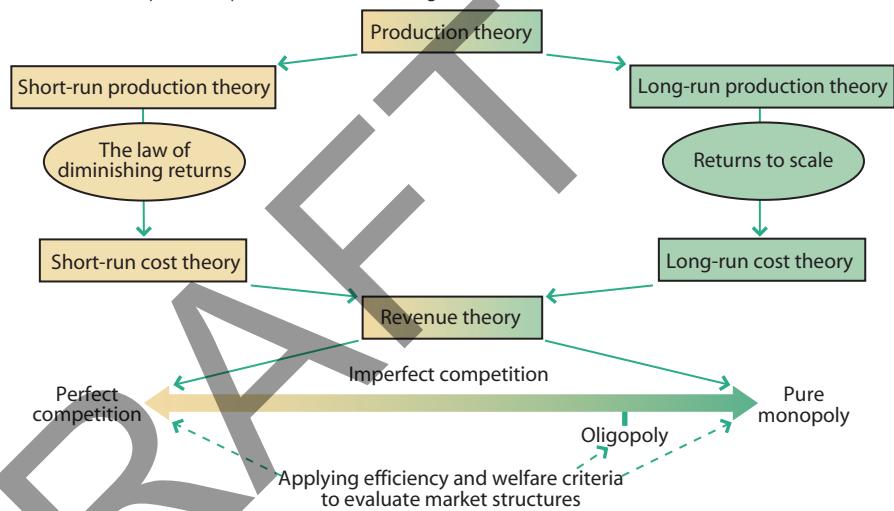
Figure 2.1 The 'building blocks' of the theory of the firm

A theme running through Chapters 2 and 3 is that it is difficult and sometimes impossible to understand properly market structures such as perfect competition and monopoly (shown in the final part of Figure 2.1) without first understanding the nature of production, costs and revenue. More narrowly, cost theory cannot be fully understood without understanding the first 'building block' in the flow chart, production theory.

In this chapter, we explain in more detail the concept, first used in Chapter 1, of 'the margin', which we use to explain production, cost and revenue curves in a more rigorous way than was the case in Book 1. The margin is one of the most important economic concepts in the A-Level economics specification, especially in microeconomics.

Figure 2.1 reminds us of the distinction explained in Book 1 between short-run and long-run production and cost theory. However, in this chapter we explain how the law of diminishing returns determines the shape of the marginal returns curve (in short-run production theory) and the marginal cost curve (in short-run cost theory). Likewise, we explain how the long-run concept of returns to scale is a major determinant of the shape of long-run cost curves.

In contrast to production and cost curves, a firm's revenue curves are determined by the competitiveness and structure of the market in which the firm sells its output. The main forms of market structure, shown in the lower part of Figure 2.1, are explained and analysed in Chapter 3.

2.1 Developing short-run production theory: the law of diminishing returns

What is a firm?

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

Before we delve further into the nature of production theory, first in the short run and then in the long run, we shall first remind you of the nature of a **firm**. A firm is a business enterprise that either produces or deals in and exchanges goods or services. Unlike non-business productive organisations, such as many charities, firms are commercial, earning revenue to cover the production costs they incur.

CASE STUDY 2.1

Ronald Coase and the nature of the firm

Way back in 1937, Professor Richard Coase, who much later in 1991 received the Nobel Prize in Economics for his insights, set out to explain why firms exist. Coase's starting point was that 'production could be carried on without any organisation at all', and could be determined solely by the price mechanism. Coase then asked 'why do firms exist?' His answer was that firms exist because they reduce transaction costs, such as search and information costs, bargaining costs, costs of keeping trade secrets, and policing and enforcement costs.

Coase then asked 'why then don't firms become bigger and bigger? Why isn't all world production carried on by a single big firm?' Coase gave two main reasons. 'First, as a firm gets larger, there may be decreasing returns to the entrepreneurial

function, that is, the costs of organising additional transactions within the firm may rise…Secondly… as the transactions which are organised increase, the entrepreneur…fails to make the best use of the factors of production.' At a certain point, the gains from economies of scale are defeated by the costs of bureaucracy.

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

Follow-up question

1 Explain the meaning of the following terms mentioned in the passage: the entrepreneurial function; transaction costs; bargaining costs; and enforcement costs.

The short run and the long run

As we explained in Book 1, Chapter 3, the short run is defined as the time period in which, in the course of production, at least one of the factors of production is fixed and cannot be varied. (By contrast, in the long run, the scale of all the factors of production can be changed.) As a simplification, we shall assume that only two inputs or factors of production are needed for production to take place — capital and labour. We shall also assume that in the short run, capital is fixed. It follows that the only way the firm can increase output in the short run is by adding more of the variable factor of production, labour, to the fixed capital.

Table 2.1 Short-run production with fixed capital

Fixed capital	Variable labour										
	0	1	2	3	4	5	6	7	8	9	10
Total returns	0	1	8	18	32	50	64	70	72	68	60
Average returns	–	1	4	6	8	10	10.7**	10	9	7.6	6
Marginal returns	1	7	10	14	18*	14	6	2	–2	–8	

* The point of diminishing marginal returns

** The point of diminishing average returns

Note: Total, average and marginal returns are often called total, average and marginal product.

For example, in Table 2.1 the 'marginal returns of labour' can be called the 'marginal product of labour'.

Table 2.1 shows what might happen in a small musical instrument workshop assembling guitars when the number of workers employed increases from 0 to 10. The first worker employed assembles 1 guitar a day, and the second and third workers respectively add 7 and 10 guitars to the workshop's total daily output. These figures measure the **marginal returns** (or marginal product) of each of the first three workers employed. The marginal returns of labour are the addition to total output brought about by adding one more worker to the labour force.

In Table 2.1, the first five workers benefit from increasing marginal returns (or increasing marginal productivity). An additional worker increases total output by more than the amount added by the previous worker. Increasing marginal returns are very likely when the labour force is small. In this situation, employing an extra worker allows the workforce to be organised more efficiently. By dividing the various tasks of production among a greater number of workers, the firm benefits from specialisation and the division of labour. Workers become better and more efficient in performing the particular tasks in which they specialise, and time is saved that otherwise would be lost as a result of workers switching between tasks.

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

Note that the impact of diminishing marginal returns does not mean that an extra worker joining the labour force is any less hardworking or motivated than his or

KEY TERMS

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

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

her predecessors. (In microeconomic theory we often assume that workers and other factors of production are completely interchangeable and homogeneous.) Any further specialisation and division of labour eventually become exhausted as more labour is added to a fixed amount of capital or machinery.
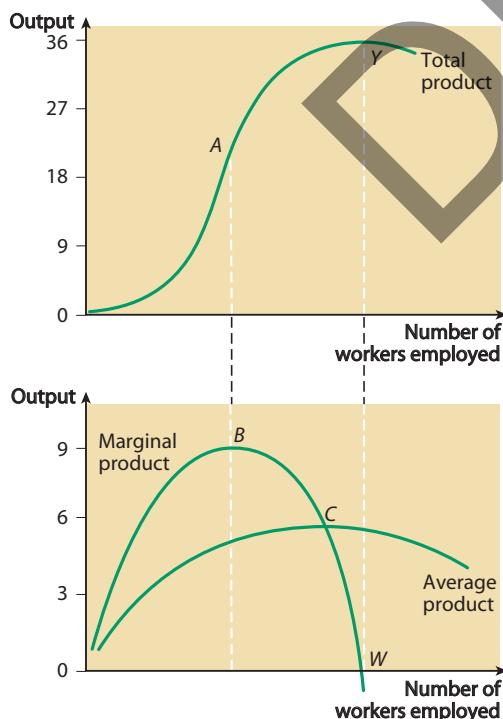
KEY TERMS

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

total returns of labour total output produced by all the workers employed by a firm.

Whereas marginal returns are the addition to total output attributable to taking on the last worker added to the labour force, the **average returns** at any level of employment are measured by dividing the total output of the labour force by the number of workers employed. The average returns of the labour force employed in the guitar workshop are shown by the middle row of data in Table 2.1. Note that in the table, the point of diminishing *average* returns occurs after the sixth worker is taken on, whereas diminishing marginal returns set in after the fifth worker is employed. The relationship between the marginal returns and the average returns of labour is illustrated in the lower panel of Figure 2.2 in the next section.

The law of diminishing returns shown on a diagram

Figure 2.2 illustrates the law of diminishing marginal returns. In the upper panel of the diagram, the law begins to operate at point *A*. Up to this point, the slope of the total product curve increases, moving from point to point up the curve. This shows the labour force benefiting from increasing marginal returns. When diminishing marginal returns set in, the **total returns** curve continues to rise as more workers are combined with capital, but the curve becomes less steep from point to point up the curve. Point *Y* shows where total returns begin to fall. Beyond this point, additional workers begin to get in the way of other workers, so the marginal returns to labour become negative.

It is important to understand that all three curves (and all three rows in Table 2.1) contain the same information, but the information is used differently in each curve (and row). The total returns curve plots the information *cumulatively*, adding the marginal returns of the last worker employed to the total returns before the worker joined the labour force. By contrast, the marginal returns curve plots the same information *non-cumulatively*, or as separate observations. Finally, at each level of employment, the average returns curve shows the total returns of the labour force divided by the number of workers employed.

Figure 2.2 Total, marginal and average returns curves

33

In the lower panel of Figure 2.2, the law of diminishing marginal productivity sets in at point *B*, at the highest point on the marginal returns curve. Before this point, increasing marginal returns are shown by the rising (or positively sloped) marginal returns curve, while beyond this point, diminishing marginal returns are depicted by the falling (or negatively sloped) marginal returns curve. Likewise, the point of diminishing average returns is located at the highest point of the average returns curve at point *C*. Finally, marginal returns become negative beyond point *W*.

QUANTITATIVE SKILLS 2.1

Worked example: diminishing returns to labour

A firm has a fixed amount of capital and land, and increases output by employing additional labour according to the schedule in Table 2.2.

Table 2.2 Diminishing returns to labour

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

1 When do diminishing marginal returns set in?
2 When do diminishing average returns set in?
3 When do diminishing total returns set in?

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

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

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

EXTENSION MATERIAL

The relationship between marginal returns and average returns

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

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

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

It is vital to understand this relationship. It does *not* state that an average will rise when a marginal is rising; nor does it state that an average will fall when a marginal falls. As we saw in Figure 2.2, marginal returns begin to fall as soon as the law of diminishing marginal returns sets in. Nevertheless, as long as marginal returns are greater than the average returns of labour, the latter continue to rise. When marginal returns exceed average returns, the average returns curve is 'pulled up', even when the marginal returns curve is falling. But when the marginal returns curve cuts through the average returns curve (at point *C* in Figure 2.2), beyond that point the average returns of labour begin to fall. The marginal returns curve cuts through the average returns curve at the latter's highest point. Beyond this point, the marginal returns curve continues to fall, and because marginal returns are less than average returns, they 'pull down' the average returns curve.

EXTENSION MATERIAL

The importance of productivity

Book 1, Chapter 3 introduced you to the very important economic concept of **productivity**, focusing in the main on **labour productivity**, or output per worker. The chapter then looked at a big problem which has adversely affected UK economic performance in recent years: the failure of labour productivity to recover from a relatively low level, compared to other countries such as Germany and the USA, in the years following the 2009 recession. This has been the called the 'productivity puzzle'. Why has the UK economy performed less well than competitor countries in increasing labour productivity?

Among the explanations of the productivity puzzle that have been put forward are: inadequate investment in new capital goods, relatively low wages in the UK economy and employers 'hoarding' rather than laying off workers in the recession, which, with depressed output, inevitably means that labour productivity falls. With regard to the latter argument, the fall in labour productivity has helped employment in the UK in the short run, but the long-run consequences of low productivity growth may be much less favourable. For further information on productivity, labour productivity and related concepts such as the UK's productivity gap, re-read Book 1, pages 55–58.

KEY TERMS

productivity output per unit of input.

labour productivity output per worker.

CASE STUDY 2.2

Japanese manufacturing methods and labour productivity

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

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

Lean manufacturing rejects the old idea of making things in huge batches, which requires holding large buffer stocks of materials and components between each stage of the production process.

Car manufacturers use lean manufacturing to eliminate waste

Now each stage of manufacturing performed in the factory is done on demand. The process eliminates waste by making only as much as is wanted at any given time; gone are the costly piles of work-in-progress that used to litter the factory floor. The change has greatly affected labour productivity.

Follow-up question

1 Explain how the changes in methods of production mentioned in the passage are likely to have affected labour productivity and costs of production within manufacturing industries.

2.2 Developing long-run production theory: returns to scale

Returns to scale

KEY TERMS

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

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

Figure 2.3, which is an extended version of Figure 3.3 in Book 1, page 61, illustrates the important distinction between returns to a variable factor of production, which occur in the short run, and **returns to scale**, which operate only in the economic long run. Suppose that a firm's fixed capital is represented by **plant** size 1 in the diagram. Initially, the firm can increase production in the short run, by moving along the horizontal arrow A, employing more variable factors of production such as labour. To escape the impact of short-run diminishing marginal returns which eventually set in, the firm may make the long-run decision to invest in a larger production plant, such as plant size 2. The movement from plant size 1 to plant size 2 is shown by the movement along the vertical arrow X in the diagram. Once plant size 2 is in operation, the firm is in a new short-run situation, able to increase output

by moving along arrow *B*. But again, the impact of short-run diminishing returns may eventually cause the firm to expand the scale of its operations to plant size 3 in the long run.
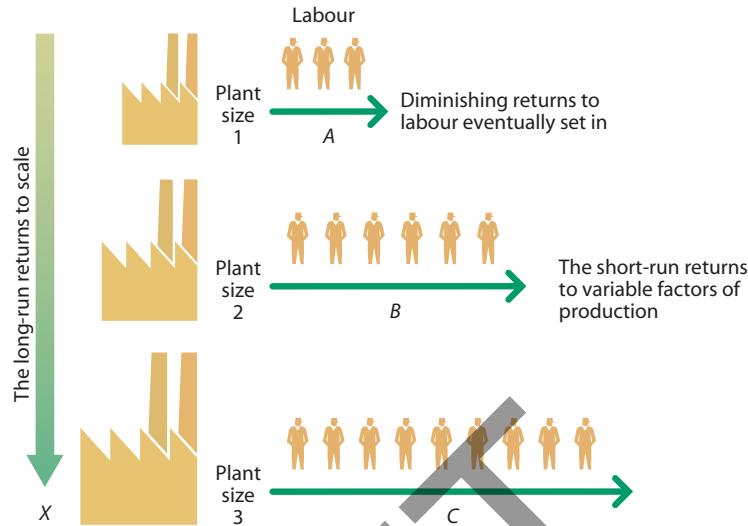
Figure 2.3 Contrasting short-run and long-run production

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

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

Economies and diseconomies of scale

Just as it is important to avoid confusing short-run returns to the variable factors of production with long-run returns to scale, so returns to scale must be distinguished from a closely related concept: economies and diseconomies of scale. Returns to scale refer to the *technical* relationship in production between *inputs* and *outputs* measured in physical units. For example, increasing returns to scale occur if a doubling of a car firm's factory size and its labour force and other factors of production enables the firm to more than double its output of cars. There is no mention of money costs of production in this example of increasing returns to scale. Returns to scale are part of long-run production theory, but economies and diseconomies of scale are part of long-run cost theory. **Economies of scale** occur when **long-run average cost** (*LRAC*) falls as output increases. **Diseconomies of scale** occur when *LRAC* rises as output increases.

KEY TERMS

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

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

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

KEY TERMS

economy of scale as output increases, long-run average cost falls.

diseconomy of scale as output increases, long-run average cost rises.

long-run average cost cost per unit of output incurred when all factors of production or inputs can be varied.

37

The link between returns to scale and economies and diseconomies of scale is that increasing returns to scale lead to falling long-run average costs or economies of scale, and likewise decreasing returns to scale bring about rising long-run average costs or diseconomies of scale. The effect of increasing returns to scale on long-run average costs can be explained in the following way: output increases faster than inputs, so if wage rates and other factor prices are the same at all levels of output, the money cost of producing a unit of input must fall. Likewise, with decreasing returns to scale, output increases at a slower rate than inputs, and the money cost of producing a unit of output rises.

There are other reasons for falling long-run average costs besides the impact on costs of increasing returns to scale. These include the effect of 'bulk buying' reducing the cost of raw materials and components.

Bringing together long-run average cost and short-run average cost

In Book 1 we described and explained various possible shapes of *LRAC* curve. In this chapter, we go a stage further by explaining the relationship between a firm's *LRAC* curve and the associated *SRATC* curves.

Figure 2.4 shows a number of short-run average total cost (*SRATC*) curves, each representing a particular firm size. In the long run, a firm can move from one short-run cost curve to another, for example from $SRATC_1$ to $SRATC_2$, with each curve associated with a different scale of capacity that is fixed in the short run. The line drawn as a tangent to the family or set of *SRATC* curves is the *LRAC* curve.
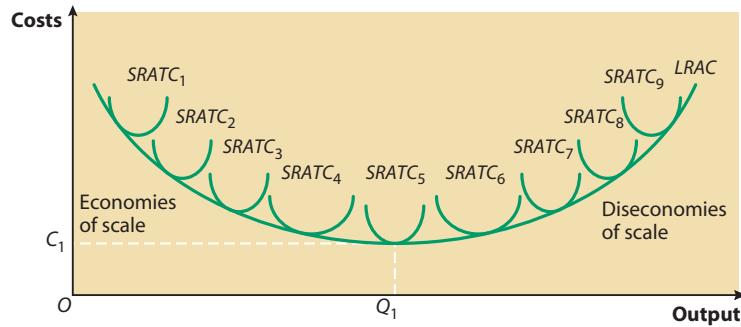
Figure 2.4 A U-shaped *LRAC* curve and its related *SRATC* curves

Optimum firm size

The size of firm at the lowest point on the firm's *LRAC* curve is known as the **optimum firm size**. In Figure 2.4, we can identify a single optimum firm size, occurring after economies of scale have been gained, but before diseconomies of scale set in. In the graph, optimum firm size is shown by the short-run cost curve $SRATC_5$, with optimum output at Q_1.

Other shapes of *LRAC* curve

The *LRAC* curve need not be symmetrically U-shaped, with a single identifiable optimum size of firm, as illustrated in Figure 2.4. Four other possibilities are depicted in Figures 2.5, 2.6, 2.7 and 2.8. Figure 2.5 is a variant of Figure 2.4, but with a horizontal section to the *LRATC* curve inserted between the sections of the curve showing economies and diseconomies of scale.
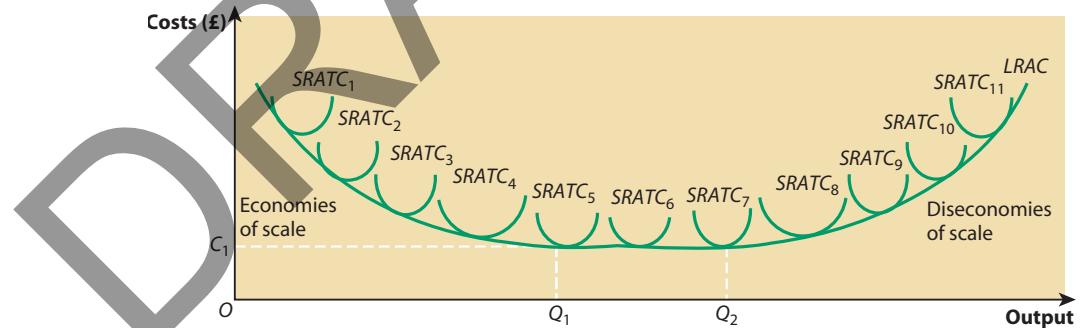
Figure 2.5 A 'three-section' long-run average cost curve

In this diagram, the *LRAC* curve comprises three sections: a downward-sloping section showing economies of scale; a horizontal mid-section; and finally, an upward-sloping section which begins when diseconomies of scale set in. With this shape of the *LRAC* curve, it is not possible to identify a single optimum firm size. Long-run average costs of production would be the same for any size of firm producing at the lowest points on $SRATC_5$, $SRATC_6$ and $SRATC_7$, between and including the levels of output Q_1 and Q_2.

Figure 2.6 illustrates an important concept in production and cost theory: **minimum efficient scale** (*MES*). *MES* is the lowest output at which long-run average costs have been reduced to the minimum level that can be achieved, which means that the firm has benefited to the full from economies of scale. In Figure 2.6, all firm sizes to the left of the $SRATC_3$ curve are below minimum efficient scale, incurring higher average costs than can be achieved at the

KEY TERMS

optimum firm size the size of firm capable of producing at the lowest average cost and thus being productively efficient.

minimum efficient scale the lowest output at which the firm is able produce at the minimum achievable *LRAC*.

lowest point on $SRATC_3$. By contrast, there would be no further reductions in long-run production costs for any firms producing levels of output above Q_1. In the diagram, the MES level of output is Q_1, with average costs minimised at C_1.
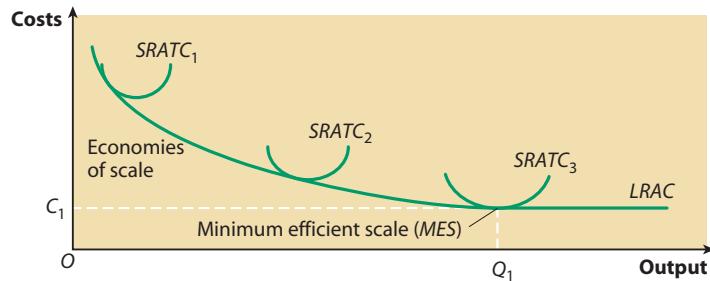

Figure 2.6 An 'L'-shaped $LRAC$ curve and minimum efficient scale (MES)

Another possibility is illustrated by Figure 2.7: an $LRAC$ curve which is horizontal throughout its length. This curve depicts a market or industry in which firms neither benefit from economies of scale nor suffer the consequences of diseconomies of scale.
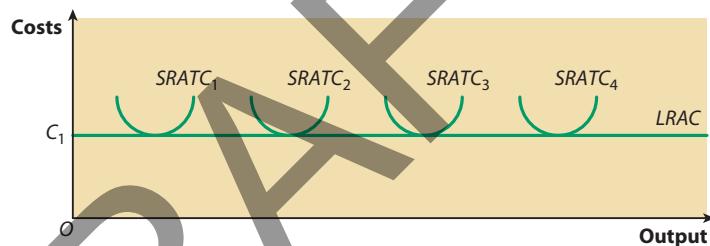
Figure 2.7 Constant long-run average costs

Figure 2.8 helps to explain why small firms are common in industries supplying services, such as those provided by hairdressers and personal trainers. The markets in which these services are provided typically possess economies of small-scale production. Diseconomies of scale may set in early in such industries, resulting in an $LRAC$ curve in which the optimum-sized firm, depicted by the short-run average cost curve $SRATC_2$, is relatively small. The MES, Q_1, is at a low level of output.
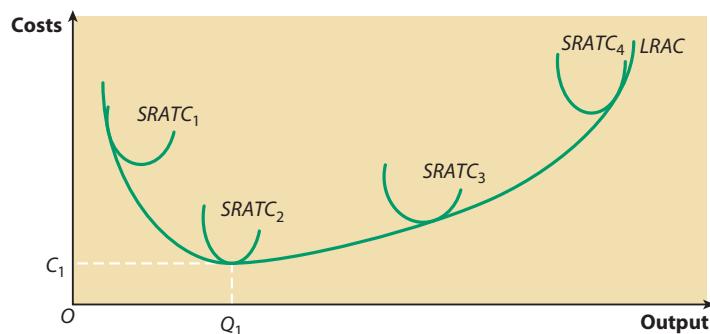
Figure 2.8 The $LRAC$ curve in an industry with economies of small-scale production

Plant-level economies of scale and firm-level economies of scale

So far, we have discussed economies and diseconomies of scale which occur when the whole of a *firm* grows in size. Sometimes, however, *firms* grow larger but without the *plants* they own and operate growing significantly in size. For this reason, it is useful to distinguish between economies of scale that occur at the level of a single plant or establishment owned by a firm and those occurring at the level of the whole firm. In recent years, continued opportunities for further firm-level economies of scale have contributed to the growth of larger firms, but expansion of plant size has been less significant.

External economies and diseconomies of scale

The scale economies and diseconomies referred to so far in this chapter have been **internal economies and diseconomies of scale**. These occur when a firm, or a plant within the firm itself, increases its scale and size. By contrast, **external economies of scale** occur when average or unit costs of production fall, not because of the growth of the firm or plant itself, but because of the growth of the industry or market of which the firm is a part. Likewise, **external diseconomies of scale** occur when average costs of production increase because of the growth of the whole industry or market. To find out more about external economies and diseconomies of scale, and also about the various types of internal economy and diseconomy of scale, you should refer back to Book 1, Chapter 3.

Scale and market structure

In real life, some markets contain just a few firms — in the extreme case of pure monopoly, just one firm. At the other extreme there are markets containing a large number of similarly sized small firms. Between these extremes are markets containing firms of a variety of sizes — some large firms but also some small firms.

The existence or non-existence of increasing returns to scale and economies of scale provide one explanation for variability in the size of firm in different market or industry structures. This section brings together some conclusions that can be drawn from Figures 2.6, 2.7 and 2.8.

- Figure 2.9, which is the same as Figure 2.6 but with a vertical line added to show the maximum size of the market, can be used to explain **natural monopoly**. Natural monopoly occurs when there is room in a market for only one firm benefiting to the full from economies of scale. In Figure 2.9 this is shown by the firm producing on the short-run average cost curve, $SRATC_3$.

> **KEY TERMS**
>
> **internal economies and diseconomies of scale**
> changes in long-run average costs of production resulting from changes in the size or scale of a firm or plant.
>
> **external economy of scale** a fall in long-run average costs of production resulting from the growth of the market or industry of which the firm is a part.
>
> **external diseconomy of scale** an increase in long-run average costs of production resulting from the growth of the market or industry of which the firm is a part.

Figure 2.9 Circumstances in which natural monopoly arises

- The horizontal *LRAC* curve in Figure 2.7 illustrates a market in which large, medium-sized and small firms or plants can coexist and compete against each other. No firm or plant gains a cost advantage, or suffers a cost disadvantage, compared to other firms or plants in the market. There are likely to be firms and plants of varying size in such a market.
- As previously noted, Figure 2.8 helps to explain why small plants or firms are common in markets or industries supplying personal services to individuals and households. Economies of small-scale production mean that the *LRAC* curve is 'skewed' to the left of the diagram. By contrast, in Figure 2.10 (below) the *LRAC* curve is 'skewed' to the right of the diagram, showing economies of large-scale production. Diseconomies of scale eventually set in, but only after substantial economies of scale have been achieved.
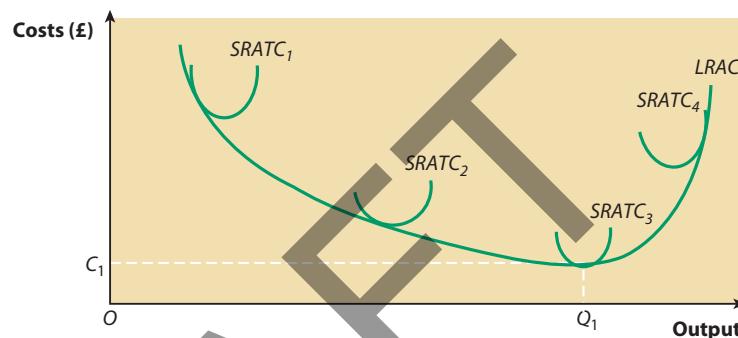

Figure 2.10 The *LRAC* curve in an industry with economies of large-scale production

There are, of course, other factors, apart from the existence or non-existence of scale economies and diseconomies, which contribute to markets containing different sizes of firm. A factor that has been becoming increasingly important in recent decades is firms 'contracting out' the provision of services, previously provided 'in house' by managers and workers employed by the firms themselves, to 'outside' suppliers of the same services. The outside suppliers range from small independent firms such as a local sandwich shop to large-scale specialist firms providing services such as catering, accountancy and ICT maintenance.

2.3 Marginal cost and marginal revenue

How a firm's short-run marginal cost curve is derived from short-run production theory

Early in this chapter, we explained the shape of a firm's marginal returns curve in relation to the impact of the short-run law of diminishing returns. In this section, we use the law of diminishing returns to explain, in the short run, the shape of a firm's **marginal cost** (*MC*) curve, before linking the *MC* curve to its short-run **average variable cost** (*AVC*) and **average fixed cost** (*AFC*) curves, which are explained in Book 1, Chapter 3, and finally its short-run **average total cost** (*ATC*) curve.
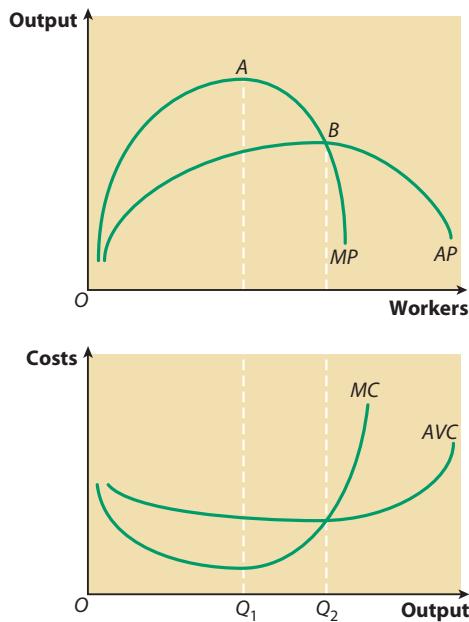
Figure 2.11 Deriving the *MC* and *AVC* curves from short-run production theory

STUDY TIP
Make sure you understand the relationship between the marginal returns curve and the short-run marginal cost curve.

Marginal cost is the extra cost a firm incurs when it produces one extra unit of output. Short-run marginal costs are determined solely by changes in variable costs of production since, by definition, in the short run, fixed costs don't change when the level of output changes. For the sake of simplicity, assuming labour is the only variable factor of production, variable costs are simply wage costs. If all workers receive the same hourly wage, total wage costs rise in exact proportion to the number of workers employed. However, if to start with the firm is benefiting from increasing marginal returns to labour, the total variable cost of production rises at a slower rate than output. This causes the marginal cost of producing an extra unit of output to fall. In Figure 2.11, the *increasing marginal returns of labour* (shown by the *positive* slope of the marginal revenue curve in the upper graph) cause *marginal costs* to fall (shown by the *negative* slope of the *MC* curve in the lower of the two graphs).

However, once the law of diminishing marginal returns has set in, short-run marginal costs rise as output increases. The wage cost of employing an extra worker is still the same, but each extra worker is now less productive than the previous worker. Total variable costs rise faster than output, so short-run marginal costs also rise. Again in Figure 2.11, the diminishing marginal returns of labour (shown by the *negative* slope of the marginal revenue curve in the right-hand side of the upper graph) cause marginal costs to rise (shown by the *positive* slope of the *MC* curve in the right-hand side of the lower graph).

Relating marginal cost to average variable cost and average total cost

Just as a firm's short-run *MC* curve is derived from the marginal returns of the variable factors of production, so the firm's average variable cost (*AVC*) curve (illustrated in the lower panel of Figure 2.11 and also in panel (a) of Figure 2.12) is explained by the average returns curve (shown in the upper panel of Figure 2.11). When increasing average returns are experienced, with the labour force *on average* becoming more efficient and productive, the *AVC* per unit of output must fall as output rises. But once diminishing *average* returns set in at point *B* in Figure 2.11, the *AVC* curve begins to rise with output.
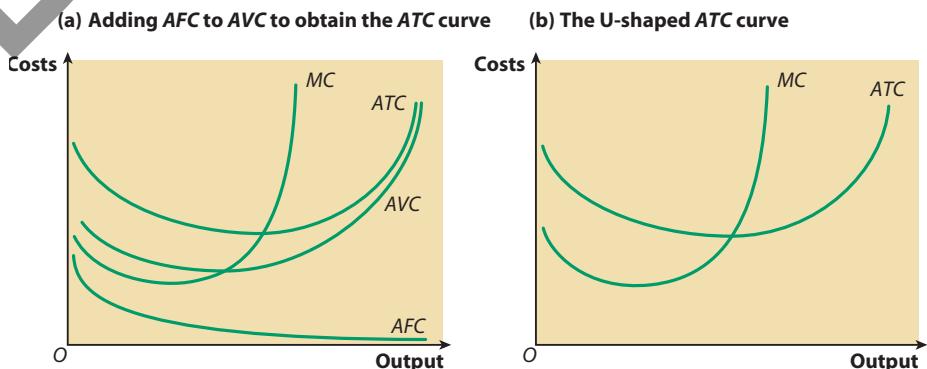
Figure 2.12 The relationships between marginal cost, average variable cost and average total cost

To see a fuller explanation of the *AVC* curve (and also the average fixed cost and short-run average total cost curves), you should re-read Book 1, pages 62–63. However, at this point we should add that the marginal cost curve cuts through from below *both* the *AVC* and *SRATC* curves at the lowest points of these curves.

This is a further example of the mathematical relationship between the marginal and average values of a variable which we explained on page XX.

QUANTITATIVE SKILLS 2.2

Worked example: calculating marginal and average variable costs

Table 2.3 shows the total cost of producing different levels of output in the short run.

Table 2.3 Total cost of different output levels

Output	Total cost (£)
0	100
1	115
2	140
3	175
4	220

Calculate:

a the marginal cost of the first unit of output
b average variable cost when output is 4 units

a Total cost increases by £15 when the first unit of output is produced, so this is the marginal cost of the first unit of output. Alternatively we can use the equation $TC_1 - TC_0 = MC_1$, which gives us the same answer, £15.

b The first row in the table indicates that total fixed costs are £100 even though output is zero. This means that total fixed costs are £100 whatever the level of output. To calculate average variable cost when 4 units are produced, we divide total variable cost (£220 – £100, which is £120) by total output, which is 4: £120/4 = £30. Thus AVC = £30.

Long-run marginal cost and long-run average cost

The mathematical relationship just described also holds in the long run, in this case between **long-run marginal cost** and **long-run average cost**. If the $LRAC$ curve is U-shaped, as in Figure 2.13, the long-run marginal cost curve cuts through the lowest point of the $LRAC$ curve. (Note that, for the sake of simplicity, $SRATC$ curves have not been included in Figure 2.13.)

KEY TERMS

long-run marginal cost addition to total cost resulting from producing one additional unit of output when all the factors of production are variable.

long-run average cost total cost of producing a particular level of output divided by the size of output when all the factors of production are variable.
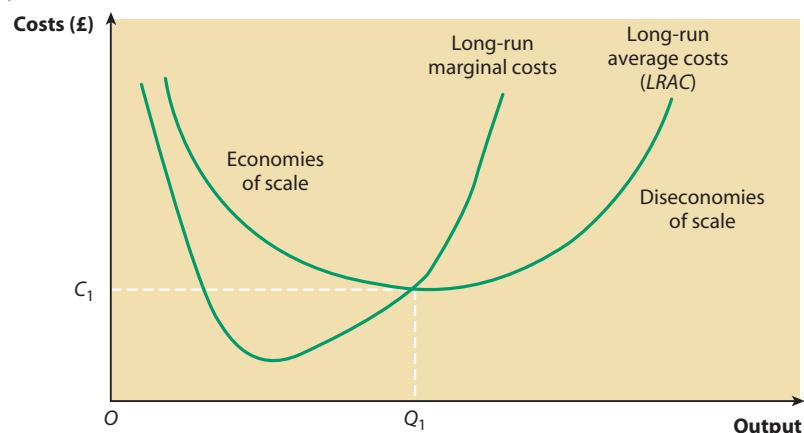
Figure 2.13 A long-run marginal cost curve cutting through a U-shaped $LRAC$ curve

Explaining the revenue curves facing firms

Book 1, page 70 introduced you to the total revenue a firm earns when selling its output, and also to a firm's average revenue. This chapter takes you deeper into revenue theory, by introducing the key concept of marginal revenue, and then by relating revenue curves to the two extreme market structures of perfect competition and monopoly. First, however, we shall summarise what you should already know about revenue from reading Book 1.

The meaning of revenue

Revenue is the money that a firm earns when selling its output. **Total revenue** (*TR*) is *all* the money a firm earns from selling the total output of a product. By contrast, at any level of output, **average revenue** (*AR*) is calculated by dividing total revenue by the size of output. Stated as an equation:

$$\text{average revenue} = \frac{\text{total revenue}}{\text{output}} \text{ or } AR = \frac{TR}{Q}$$

Marginal revenue (*MR*) is the addition to total revenue resulting from the sale of one more unit of output. Stated as an equation:

$$\text{marginal revenue} = \frac{\Delta \text{ total revenue}}{\Delta \text{ output}} \text{ or } MR = \frac{\Delta \text{ } TR}{\Delta \text{ } Q}$$

where Δ is the symbol used to indicate the changes in total revenue and the change in total output.

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

average revenue total revenue divided by output.

marginal revenue addition to total revenue resulting from the sale of one more unit of the product.

STUDY TIP
The Greek delta symbol Δ is used by mathematicians as the symbol for a change in the value of a variable over a range of observations. The word 'marginal' means the change in the value of a variable when there is one more unit of the variable, so Δ is the symbol that indicates this change. It is used in the formulae for marginal product and marginal cost, as well as for marginal revenue.

How the competitiveness of a market structure affects a firm's revenue curves

The nature of a firm's revenue curves depends on the competitiveness of the market structure in which the firm sells its output. The final row of Figure 2.1 at the start of this chapter sets out the four market structures that you need to understand. These are perfect competition, monopolistic competition, oligopoly and monopoly. However, we shall leave until Chapter 3 the imperfectly competitive market structures of monopolistic competition and

oligopoly. This means that this chapter considers only the two extreme market structures of perfect competition and monopoly.

In the latter part of Chapter 3, we shall also investigate the dynamics of competition and the competitive market processes existent in real-world markets. In this chapter, however, and in the early parts of Chapter 3, we consider solely the type of competition known as *price competition*. This focuses on the price or prices that firms charge, in pursuit of a single assumed business objective: profit maximisation. (Other types of competition include quality competition and after-sales service competition.) In this chapter we look at revenue curves, first in perfect competition and then in monopoly.

Average revenue and marginal revenue in perfect competition

A perfectly competitive market is defined by a number of conditions or characteristics that the market must possess. These conditions, which we shall revisit in Chapter 3, are:

- a very large number of buyers and sellers
- all buyers and sellers possess perfect information about what is going on in the market
- consumers can buy as much as they wish to purchase and firms can sell as much as they wish to supply at the ruling market price set in the market as a whole
- an individual consumer or supplier cannot affect the ruling market price through its own actions
- an identical, uniform or homogeneous product
- no barriers to entry into, or exit from, the market in the long run

Taken together, the six listed conditions tell us that a perfectly competitive firm, which is depicted in panel (a) of Figure 2.14, faces a perfectly elastic demand curve for its product. (Figure 2.14 is the same as Figure 4.2 in Book 1, page 77.) The demand curve facing the firm is located at the ruling market price, P_1, which itself is determined through the interaction of market demand and market supply in panel (b) of the diagram. Note that the horizontal axis in the panel (b) diagram shows millions of units of output being produced. This is because panel (b) depicts the whole market, comprising very large numbers of both consumers and firms. In equilibrium, where market demand equals market supply, the ruling market price is P_1, and the equilibrium quantity is Q_1 millions of units. In panel (a), the horizontal axis is labelled 'hundreds', to reflect the fact that in perfect competition a single firm is only a tiny part of the total market,

The assumption that a perfectly competitive firm can sell whatever quantity it wishes at the ruling market price P_1, but that it cannot influence the ruling market price by its own action, means that all firms in perfectly competitive markets are passive **price-takers**.

The labels 'No sales' and 'No sense' placed on Figure 2.14(a), respectively above and below the price line P_1, help to explain why a perfectly competitive firm is a price-taker. 'No sales' indicates that if the firm raises its selling price above the ruling market price, customers desert the firm to buy the identical products (perfect substitutes) available from other firms at the ruling market price. 'No sense' refers to the fact that, although a perfectly competitive firm *could* sell its output below the price P_1, doing so is inconsistent with the

KEY TERM

price-taker a firm which is so small that it has to accept the ruling market price. If the firm raises its price, it loses all its sales; if it cuts its price, it gains no advantage.

profit-maximising objective. No extra sales can result, so selling below the ruling market price inevitably reduces both total sales revenue and therefore profit, given the fact that the firm can sell any quantity it wants at the ruling market price.
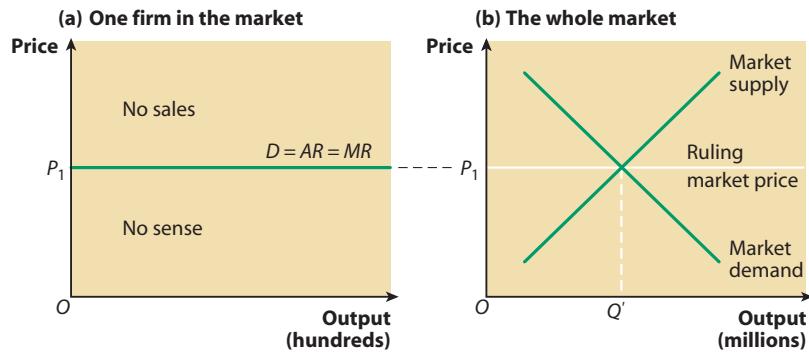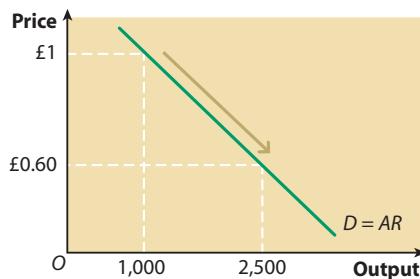

Figure 2.14 Deriving a perfectly competitive firm's average and marginal revenue curves

The horizontal price line facing a perfectly competitive firm is also the firm's average revenue (*AR*) curve and its marginal revenue (*MR*) curve. Suppose, for example, that the firm sells 100 units of a good, with each unit of the good priced at £1. The firm's total sales revenue (*TR*) is obviously £100. The horizontal price line is also the demand curve, depicted by the letter *D* on the graph, facing each firm in the market. You should note that the curve is perfectly elastic. This results from the fact that the goods produced by all the firms in the market, being uniform or homogeneous, are perfect substitutes for each other.

> **STUDY TIP**
> Note that Figure 2.14 contains two demand curves. On the one hand, the market demand curve, drawn in panel (b) of the diagram, slopes downward, reflecting the fact that in the market as a whole a fall in price will lead to an increase in the quantity sold; with a straight-line market demand curve, the price elasticity of demand falls from point to point moving down the demand curve. Goods produced in other markets are partial, though not perfect, substitutes for goods produced in this market. By contrast, the demand curve for the output of any one firm *within* the market is perfectly elastic, and hence horizontal, because other firms *within* the market produce identical goods which are perfect substitutes for the goods produced by any one firm in this market.

Average revenue and marginal revenue in monopoly

It is worth repeating that the demand curve facing a perfectly competitive firm, besides being located at the ruling market price, is also the firm's *AR* curve and its *MR* curve. By contrast, the demand curve for a monopolist's output is the monopolist's *AR* curve, but it is *not* the monopolist's *MR* curve.

To understand why the market demand curve is the monopolist's average revenue (*AR*) curve, consider Figure 2.15, which shows two prices, £1 and £0.60, which can be charged by a monopolist for the good it produces.
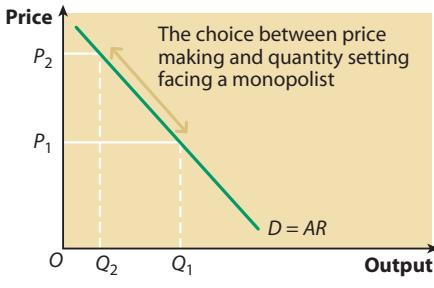

Figure 2.15 Price equalling average revenue (*AR*) in monopoly

KEY TERMS

price-maker when a firm faces a downward-sloping demand curve for its product, it possesses the market power to set the price at which it sells the product.
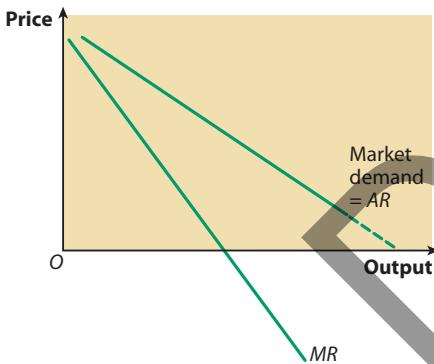
quantity-setter when a firm faces a downward-sloping demand curve for its product, it possesses the market power to set the quantity of the good it wishes to sell.

Figure 2.16 The choice between price making and quantity setting facing a monopolist

Figure 2.17 Monopoly average revenue (*AR*) and marginal revenue (*MR*) curves

Figure 2.18 Explaining a monopolist's marginal revenue (*MR*) curve

At a price of £1, 1,000 units are demanded. At this price, the monopolist's total revenue is £1000. Average revenue, or total revenue divided by output ($TR \div Q$), is £1, which is of course the same as price. This is the case at all prices: the price charged for all units of the good and average revenue are always the same. For example, if the monopolist sets the price at £0.60, 2,500 units of the good are demanded; total sales revenue is £1,500 and average revenue ($TR \div Q$) is £0.60. The demand-sloping market demand curve facing the monopolist is therefore the firm's average revenue (*AR*) curve.
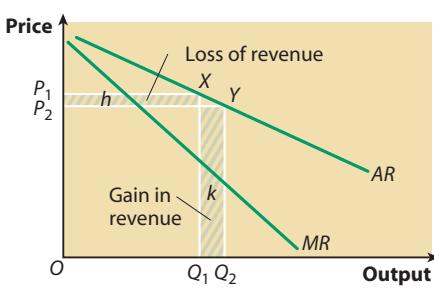
The downward-sloping *AR* curve can affect the monopoly in two different ways. If the monopolist is a **price-maker**, choosing to set the price at which the product is sold, the demand curve dictates the maximum output that can be sold at this price. For example, if the price is set at P_1 in Figure 2.16, the maximum quantity that can be sold at this price is Q_1. But if the monopolist cuts the price it charges to P_2, sales increase to Q_2. Alternatively, if the monopolist is a **quantity-setter** rather than a price-maker, the demand curve dictates the maximum price at which a chosen quantity of the good can be sold. If the monopolist wants to sell Q_2, the market demand curve shows that the maximum price at which this quantity can be sold is P_2. To summarise, if the monopolist sets the price, the market demand curve dictates the maximum quantity the firm can sell. Conversely, if the monopolist sets the quantity, the market demand curve determines the maximum price the firm can charge. However, for any one good it produces, a firm cannot be a price-maker and a quantity-setter at the same time.

However, to understand why marginal revenue and average revenue are *not* the same in monopoly, you must remember that when the *marginal* value of a variable is less than the *average* value of the variable, the *average* value falls.

Because the market demand curve or average revenue curve falls as output increases, the monopolist's marginal revenue curve *must* be below its average revenue curve. Figure 2.17 shows the relationship between the *AR* and the *MR* curves. You should see, however, that the *MR* curve is not only below the *AR* curve — it has also been drawn twice as steep. This is always the case whenever the *AR* curve is a downward-sloping straight line.

The relationship between *AR* and *MR* curves is illustrated in Figure 2.18. The monopolist initially charges a price of P_1 and sells the level of output Q_1. However, to increase sales by an extra unit to Q_2, the downward-sloping *AR* curve forces the monopolist to reduce the selling price to P_2. This reduces the price at which *all* units of output are sold. Total sales revenue increases by the area k in Figure 2.18, but decreases by the area h. Areas k and h respectively show the revenue gain (namely, the extra unit sold multiplied by price P_2) and the revenue loss resulting from the fact that, in order to sell more, the price has to be reduced for *all* units of output, not just the extra unit sold. Marginal revenue, which is the revenue gain minus the revenue loss (or $k - h$), must be less than price or average revenue (area k).

EXTENSION MATERIAL

Elasticity and revenue curves

We mentioned earlier in the chapter that the horizontal price line facing a perfectly competitive firm is also the perfectly elastic demand curve for the firm's output. The explanation for this lies in the word 'substitutability'. When studying elasticity, you learnt that the availability of substitutes is the main determinant of price elasticity of demand. In perfect competition, because of the assumptions of a uniform product and perfect information, the output of every other firm in the market is a perfect substitute for the firm's own product. If the firm tries to raise its price above the ruling market price, it loses all its customers.

In monopoly, by contrast, providing the demand curve is a straight line as well as downward sloping, price elasticity of demand falls moving down the demand curve. Demand for the monopolist's output is elastic in the top half of the curve, falling to be unit elastic exactly half way down the curve, and inelastic in the bottom half of the curve. This is shown in

Figure 2.19. Demand is elastic between *A* and *B*, unit elastic at *B*, and inelastic between *B* and *C*.

We shall revisit the significance of elasticity in the next chapter, when comparing profit maximisation with revenue maximisation.
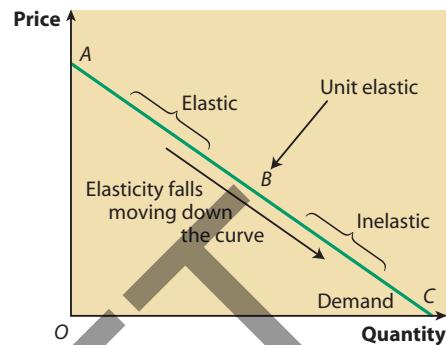

Figure 2.19 Price elasticity of demand and a monopolist's demand or average revenue (*AR*) curve

2.4 Profit

Defining profit

We mentioned in Book 1, Chapter 3 that students often confuse profit and revenue, mistakenly believing that the two terms have the same meaning. In fact, profit and revenue are different. Revenue has already been explained in some depth in the previous sections of this chapter, and profit has been briefly mentioned on a number of occasions. **Profit** is the difference between the sales revenue the firm receives when selling the goods or services it produces and the costs of producing the goods.

    total profit = total revenue − total costs

KEY TERM
profit the difference between total sales revenue and total cost of production.

KEY TERM
profit maximisation occurs at the level of output at which total profit is greatest.

The difference between normal and abnormal (supernormal) profit

We have mentioned on several occasions in this chapter, and also in Book 1, page 30, that economists often assume that firms have a single business objective: **profit maximisation**. This means producing the level of output

at which profit (revenue minus costs) is greatest. (Firms may also have other objectives, such as survival, growth and increasing their market share.)

In the next chapter, we shall explain how profit maximisation is achieved in the different market structures of perfect competition, monopoly, monopolistic competition and oligopoly. When explaining profit maximisation, we shall apply two profit concepts, used frequently by economists undertaking microeconomic analysis, but rarely used outside the field of microeconomic theory. These are *normal profit* and *abnormal profit*. (Abnormal profit is also called *supernormal profit* and *above-normal profit*.)

Normal profit

Normal profit is the minimum level of profit necessary to keep incumbent firms in the market, rewarding the time, decision making and entrepreneurial risk taking 'invested' into production. However, the normal profit made by incumbent firms, or firms already established in the market, is insufficient to attract new firms into the market. Economists treat normal profit as an opportunity cost, which they include in firms' average cost curves. In the long run, firms unable to make normal profit leave the market. Normal profit varies from one industry to another, depending on the risks facing firms.

Abnormal profit

Abnormal profit, or supernormal profit, is extra profit over and above normal profit. In the long run, and in the absence of entry barriers, abnormal profit performs the important economic function of attracting new firms into the market.

> **KEY TERMS**
>
> **normal profit** the minimum profit a firm must make to stay in business, which is, however, insufficient to attract new firms into the market.
>
> **abnormal profit** profit over and above normal profit. Also known as supernormal profit and above-normal profit.

> **STUDY TIP**
>
> Avoid confusing normal profit with another abstract microeconomic term: normal good. You came across normal goods when studying demand theory in Book 1. A normal good is a good for which demand increases as income increases.

QUANTITATIVE SKILLS 2.3

Worked example: calculating revenue and profit

Table 2.4 provides information about the short-run output, costs and revenue of a firm.

Table 2.4 Short-run output, costs and revenue

Output per week	Total revenue (£000s)	Total cost (£000s)
0	0	10
1	20	14
2	38	19
3	54	28
4	68	44
5	80	80
6	90	93

From the information in the table, calculate:

a marginal revenue when output per week increases from 4 to 5 units

b the level of output at which the firm would make normal profit but not abnormal profit

c the profit-maximising level of output per week

a The marginal revenue, which is the change in the total revenue, is £80,000 – £68,000, which is £12,000.

b Assuming that normal profit is being treated as a cost of production, the firm makes normal profit, but not abnormal profit, when total revenue equals total cost. This is at a level of output of 5 units per week.

c Profits are maximised at the level of output at which (*TR – TC*) is greatest. This is at a level of output of 3 units per week, when total profit equals £54,000 – £28,000, which is £26,000.

> **TEST YOURSELF 2.3**
> Using the information in Table 2.4, draw on a piece of graph paper the firm's average revenue and marginal revenue curves.

The role of profit in a market economy

Profit performs a number of roles in a market economy. These include the creation of business, worker and shareholder incentives. Profit also influences the allocation of resources, it is an efficiency indicator, and it is a reward for innovation and for risk taking. Finally, profits also provide an important source of business finance.

The creation of business incentives

As we have noted on several occasions both in this book and in Book 1, traditional or orthodox microeconomic theory assumes that profit maximisation is the most important business objective. Not only do rising profits, and the hope of higher profits in the future, provide the incentive for managers within a firm to work harder to make the business even more profitable, but also they create incentives for other firms to enter the market. Abnormal profit acts as a 'magnet' attracting new entrants into a market or industry. If market entry is easy and/or relatively costless, new firms joining the market should lead to an increase in market supply. We shall explain in the next chapter how the entry of new firms triggers a process which actually reduces both abnormal profit and prices, with the latter benefiting consumers. We shall also explain how economic efficiency and economic welfare may be promoted through this process.

However, when entry barriers are high and monopoly or highly imperfect competition exists in a market, profit may simply reward inefficient producers. This is a form of market failure in which the 'producer is king' rather than the consumer, and in which 'producer sovereignty' rather than 'consumer sovereignty' exists.

The creation of worker incentives

Some companies use profit-related pay and performance-related pay to increase worker motivation, in the hope that workers will work harder and share the objectives of the business's managers and owners. This can, however, be counter-productive, if ordinary workers see higher management and company directors enjoying huge profit-related bonuses, while they receive a pittance.

The creation of shareholder incentives

High profit generally leads to high dividends or distributed profit being paid out to shareholders who own companies. This creates an incentive for more people to want to buy the company's shares. As a result, the company's share price rises, which makes it cheaper and easier for a business to raise finance.

Profits and resource allocation

High profits made by incumbent firms in a market create incentives for new producers to enter the market and for existing firms to supply more of a good or service. Likewise, loss making, or perhaps a failure to make above-normal profits, create incentives for firms to leave markets and to deploy their resources in more profitable markets.

Profit and economic efficiency

Except when monopolies make large profits by exploiting their consumers, profit can be an indicator of economic efficiency. Large profits might mean that firms have succeeded in eliminating unnecessary costs of production and are also using the most efficient production processes.

Profit as a reward for innovation and risk taking

As we explain later in this chapter, innovation is an improvement on something that has already been invented, which thus turns the results of invention into a useful product. If entrepreneurs believe that innovation can result in high profits in the future, the incentive to innovate increases. As we can never be sure of future profits, risks are involved. However, successful risk taking leads to high profits.

Profit as a source of business finance

Instead of being distributed to the business's owners as a form of income, profit can be retained within the business. Retained profits are perhaps the most important source of finance for firms undertaking investment projects. High profits also make it easier and cheaper for firms to use borrowed funds as an important source of business finance.

Profit sends out a signal about the health of the economy

The profit made by businesses throughout the economy can send out an important signal about the health of the macroeconomy. Rising profit may reflect improvements in supply-side performance — for example, higher productivity or lower costs resulting from innovation.

CASE STUDY 2.3

The John Lewis economy

John Lewis PLC is a highly successful retailing company in the UK which not only shares its profits with all its workers, but also makes them part-owners of the John Lewis Partnership. From time to time, leading British politicians climb aboard the John Lewis bandwagon, extolling the virtues of the British retailer and exhorting other companies to introduce co-ownership and profit sharing. Here is an extract from a speech made by Nick Clegg, the then deputy prime minister, on 16 January 2012 at an event hosted by the City of London.

We need more individuals to have a real stake in their firms. More of a John Lewis economy, if you like. What many people don't realise about employee ownership is that it is a hugely underused tool in unlocking growth. I don't value employee ownership because I believe it is somehow 'nicer' — a more pleasant alternative to the rest of the corporate world. Those are lazy stereotypes. Firms that have engaged employees, who own a chunk of their company, are just as dynamic, just as savvy, as their competitors. In fact, they often perform better: lower absenteeism, less staff turnover, lower production costs. In general, they have higher productivity and higher wages. They weathered the economic downturn better than other companies.

Is employee ownership a panacea? No. Does it guarantee a company will thrive? Of course not. But the evidence and success stories cannot be ignored, and we have to tap this well if we are serious about growth. The 1980s was the decade of share ownership. I want this to be the decade of employee share ownership.

Follow-up question

1 Outline one advantage and one disadvantage of profit sharing and co-ownership for the United Kingdom economy.

2.5 Technological change

Most people have a general idea of what technology means, but they nevertheless find it difficult to give the term a precise definition. Internet search engines, being prone to long-winded explanations rather than to short, snappy definitions, often don't help in this respect. Here, however, is one snappy definition: whereas science is concerned with *how* and *why* things happen, technology focuses on *making things happen*. Thus, technology is knowledge put to practical use to solve problems facing human societies.

Technological change, by contrast, involves improving existing technologies and the development of completely new technologies, both to improve existing *products* and the *processes* involved in making the products, and to develop completely new products and processes. In the economic sphere, this leads to the development of completely new markets, to changes in market structure, and also to the destruction of existing markets.

As an aside, the word 'technology' is often associated with technical progress. However, the term 'technical progress' has two rather different meanings. On the one hand, in a normative or value-judgement context, 'technical progress' implies that technological change is fundamentally about increasing economic welfare and making people happier. For example, although the development and use of the motor car has several important drawbacks, such as the harm resulting from road accidents and environmental pollution, for the most part, through making it much easier for most people to travel, cars and buses have significantly improved human welfare. But in a narrower sense, unrelated to welfare considerations, 'technical progress' means applying scientific and engineering knowledge, as it develops, to produce goods which are more efficient and work better, regardless of whether these are good for society. In this narrower meaning, technical progress includes the development of distinctly harmful goods such as chemical weapons, which, when used, have a devastating effect on human welfare.

The difference between invention and innovation

Invention refers to advancements in pure science, whereas **innovation** is the application of the new knowledge created by invention to production. The American entrepreneur, Tom Grasty, distinguishes between the two concepts in the following way:

> In its purest sense, invention can be defined as the creation of a product or introduction of a process for the first time. Innovation, on the other hand, occurs if someone improves on or makes a significant contribution to an existing product, process or service. Consider the microprocessor. Someone invented the microprocessor. But by itself, the microprocessor was nothing more than another piece on the circuit board. It's what was done with that piece — the hundreds of thousands of products, processes and services that evolved from the invention of the microprocessor — that required innovation.
>
> If ever there were a poster child for innovation it would be former Apple CEO Steve Jobs. And when people talk about innovation, Jobs' iPod is

cited as an example of innovation at its best. But let's take a step back for a minute. The iPod wasn't the first portable music device (Sony popularized the 'music anywhere, anytime' concept 22 years earlier with the Walkman); the iPod wasn't the first device that put hundreds of songs in your pocket (dozens of manufacturers had MP3 devices on the market when the iPod was released in 2001); and Apple was actually late to the party when it came to providing an online music-sharing platform (Napster, Grokster and Kazaa all preceded iTunes).

So, given those sobering facts, is the iPod's distinction as a defining example of innovation warranted? Absolutely. What made the iPod and the music ecosystem it engendered innovative wasn't that it was the first portable music device. It wasn't that it was the first MP3 player. And it wasn't that it was the first company to make thousands of songs immediately available to millions of users. What made Apple innovative was that it combined all of these elements — design, ergonomics and ease of use — in a single device, and then tied it directly into a platform that effortlessly kept that device updated with music.

Apple invented nothing. Its innovation was creating an easy-to-use ecosystem that unified music discovery, delivery and device. And, in the process, they revolutionized the music industry.

The original iPod, which was launched in 2001 — invention or innovation?

How technological change affects methods of production, productivity, efficiency and firms' costs of production

Through its diffusion into the economy, technological change affects methods of production, productivity, efficiency and firms' costs of production. We shall now look at each of these in turn.

Methods of production

Throughout human history, technological change has affected methods of production. As far back as the stone age, the bronze age and the iron age, as the names themselves indicate, different materials were used to create early forms of tool used by humankind. These eras covered scores of thousands of years. Moving much closer to the present day, the eighteenth century saw the onset of agricultural and industrial revolutions in which technological change greatly affected methods of agricultural and industrial production. Over the decades and centuries that followed, and right up to the present day, agricultural output greatly increased with the development of new seeds and the breeding of modern farm animals, the mechanisation of production — for example, through the use of combine harvesters — and the application of chemical fertilizers.

At the beginning of the industrial revolution in the eighteenth century, manufacturing moved away from craft and cottage industry, to factory production. Much more energy was now used in the course of production, so manufacturing moved to parts of the country where first water power and then steam power, fired by the burning of coal, was in plentiful supply. In the nineteenth century, the 'new' iron age of the eighteenth century in which cast iron had become perhaps the main industrial raw material, gave way to a 'steel age', when improvements in smelting technology enabled steel, which is strong yet malleable, to replace iron in much of modern manufacturing.

The steel age of the mid-nineteenth century was accompanied by a railway-building age, in which railways replaced the rather cumbersome system of eighteenth-century canals to enable the development of a modern transport system necessary for the delivery of goods to markets.

The mid-twentieth century witnessed the growing use of automobiles. The 'automobile age' really got going in the USA in the 1920s, when modern roads were built across the North American continent and mass production allowed affordable cars to be bought by much of the US population. To bring us right up to the present day, we are now living in the 'computer age'. Computers, which were first developed in the 1940s, are now widely used in manufacturing (for example, when computer-controlled robots build cars), in distribution (for example, in the online sale of books by Amazon), and as consumer goods in themselves. And hidden within many goods that are not themselves computers, such as washing machines and cars, are microprocessors that control how the good functions.

Closely allied to the changes in production just described has been the change in recent decades from mechanised to automated production. As a simplification, **mechanisation** means that human beings operate the machines that are used to produce goods. **Automation**, by contrast, means that machines operate other machines — for example, a computer-controlled robot operating a welding tool to weld together the body panels of a car. Both mechanisation and automation have often been accompanied by assembly-line production, allegedly first introduced by Henry Ford in 1908.

Productivity

Earlier in this chapter on page XX, we reminded you of the meaning of *productivity*, a key concept which we explained in some detail in Book 1. We defined productivity as output per unit of input, though it can also be considered as output per unit of time. We also said that when economists talk about productivity, they usually mean *labour productivity*, which is output per worker.

Technological change generally increases labour productivity. This has usually been the case following the introduction of both mechanised and automated production methods. However, in the case of automation and the use of computers in production, there have been several well-publicised examples in organisations such as the National Health Service of very expensive computer systems that have failed to work properly and which, in extreme cases, have had to be scrapped. In these cases, labour productivity may fall rather than increase, at least until the system can be made to work properly.

CASE STUDY 2.4

The failure to produce paperless records of patient care in the NHS

On 18 September 2013, the Parliamentary Public Accounts Committee published a report on the Dismantled National Programme for IT in the NHS. Although officially 'dismantled', the National Programme continues in the form of separate component programmes which are still racking up big costs.

Launched in 2002, the National Programme was designed to reform the way that the NHS in England uses information. While some parts of the National Programme were delivered successfully, other important elements encountered significant difficulties. In particular, there were delays in developing and deploying the detailed care records systems. Following three reports on the National Programme by both the National Audit Office and the Public Accounts Committee, and a review by the Major Projects Authority, the government announced in September 2011 that it would dismantle the National Programme but keep the component parts in place with separate management and accountability structures.

The public purse is continuing to pay the price for failures by the department and its contractors. The department's original contracts totalled £3.1 billion for the delivery of care records systems to 220 trusts in the north, midlands and east. The full

cost of the National Programme is still not certain. The department's most recent statement reported a total forecast cost of £9.8 billion. However, this figure did not include potential future costs.

The benefits to date from the National Programme are extremely disappointing. The department's benefits statement reported estimated benefits to March 2012 of £3.7 billion, just half of the costs incurred to this point. The benefits include financial savings, efficiency gains and wider benefits to society (for example, where patients spend less time chasing referrals). However, two-thirds of the £10.7 billion of total forecast benefits were still to be realised in March 2012.

After the sorry history of the National Programme, the Public Accounts Committee was sceptical that the department could deliver its vision of a paperless NHS by 2018. Making the NHS paperless will involve further significant investment in IT and business transformation.

Follow-up questions

1 Research on the internet to find out what has happened to the NHS's plan to introduce paperless patient records in the period since September 2013.
2 Find out about and investigate another example of a 'computer systems disaster' in either the public sector or the private sector.

Efficiency

As we explain in Chapter 3, economists recognise a number of types of economic efficiency. Two of these are productive efficiency and dynamic efficiency. **Productive efficiency** centres on minimising average costs of production. **Dynamic efficiency** measures the extent to which productive efficiency increases over time, in the economic long run. Dynamic efficiency also results from improvements in products and services, innovation and the process of creative destruction.

Technological change generally improves both productive efficiency and dynamic efficiency. As a general rule — though there are exceptions, one of which is illustrated by Case Study 2.4 — technological change leads to improvements in both productive and dynamic efficiency. By increasing productivity, over time technological changes shift downward both short-run and long-run cost curves, thereby improving both productive and dynamic efficiency.

KEY TERMS

productive efficiency involves minimising the average costs of production.

dynamic efficiency measures improvements in productive efficiency that occur in the long run over time.

57

Costs of production

It follows from what we have written about technological change generally improving both productivity and efficiency that it also reduces costs of production, in the short run but especially in the long run.

Technological change and the development of new products and new markets, and the destruction of existing markets

A theme running through this section on technological change is that, particularly in recent years, technological change has been highly significant in the development of new products and new markets, and the destruction of existing markets. To explain this further, it is useful to introduce the concepts of *disruptive* innovation and *sustaining* innovation.

A disruptive innovation is an innovation that helps create a new market, but in so doing eventually disrupts an existing market over a few years or decades, thereby displacing an earlier technology. Disruptive innovation often improves a product or service in ways that the market did not initially expect. It creates new goods or services for a different set of consumers in a new market which competes with the established market. By doing so, it eventually lowers prices in the existing market. By contrast, a sustaining innovation does not create new markets but develops existing markets, enabling firms within them to offer better value and often to compete against each other, sustaining improvements.

According to Harvard University business professors Joseph L. Bower and Clayton M. Christensen, one of the most consistent patterns in business is the failure of leading companies to stay at the top of their industries when technologies or markets change. Writing back in the 1990s, Bower and Christensen gave the example of Xerox, the US company, which at the time had dominated the photocopier market, losing market share to the Japanese company, Canon, in the small photocopier market.

Bower and Christensen ask why it is that companies like Xerox invest aggressively — and successfully — in the technologies necessary to retain their current customers, but then fail to make certain other technological investments that customers of the future will demand? The explanation they offer is that companies that dominate an existing technology are in danger, as disruptive innovation occurs, of remaining too close to their existing body of customers. All too often existing customers reject the goods produced by a new technology because it does not address their needs as effectively as a company's current products. The large photocopying centres that represented the core of Xerox's customer base at first had no use for small, slow table-top copiers produced by Xerox's new technology. Result: Canon stepped in, quickened the speed of the copiers, and took over the market.

CASE STUDY 2.5

Kodak faces an uncertain future

In September 2013, the American camera company Kodak emerged from the bankruptcy it had been in for nearly two years. Since 2000, demand for Kodak's most successful product, camera film, which once ranked among the most profitable consumer products ever invented, had been in rapid decline. To make matters worse, Kodak's management had grossly underestimated the speed of the collapse.

This was all due to the development of digital cameras. Global sales of traditional photographic film and paper had dropped like a stone. Up to that point, the boxes of film that Kodak produced had been highly profitable. New entrants were deterred by the high costs of entry to this capital-intensive market and Kodak enjoyed profit margins that might have been as high as 50%.

But in the first few months of this century, a technology change began that was to wreck one of the most lucrative business models of the last century and threaten Kodak's very existence. In 1999, only 5% of new cameras sold in the USA were digital. By the end of 2000, this had changed dramatically. By 2003, now being accused of ignoring the revolution until it was too late, Kodak cut tens of thousands of jobs at its capital-intensive film factories and announced a new digital strategy.

Disruptive digital technology has caused the crash of many business models since 1999 — but few quite so rapid as the fate that befell Kodak. Within a matter of months, the once hugely profitable camera film market had given way to the surge of digital cameras.

Follow-up questions

1 Research what has happened to Kodak since the company emerged from bankruptcy in 2013.
2 In what way have smartphones, which were first marketed in 2007, affected the market for digital cameras?

The influence of technological change on the structure of markets

The Kodak case study you have just read provides a good example of the influence of technological change on the structure of markets. The case study describes how, when cameras transmitted the images they photographed onto chemical film, very high entry barriers into the chemical film market led to Kodak's domination of the market. By contrast, entry into the digital camera market is relatively easy. Hence the camera film market, dominated by Kodak, was close to a monopoly, whereas the digital camera market is closer to a much more competitive form of market, **monopolistic competition**, which we explain in the next chapter.

Technological change does not always, however, lead to more competitive market structures. In some industries, technological change has led to outcomes in which very large firms dominate. Sometimes technological change leads to capital indivisibilities, which occur when very large quantities of capital equipment are required for one unit of a good to be produced. A good example is the jumbo jet industry. The technological change which enabled very large jet airliners to be produced led to an outcome in which, in the western world, the American Boeing Corporation and the European Airbus consortium are the only two jumbo jet manufacturers. In the next chapter, we shall call this situation a **duopoly**.

KEY TERMS

monopolistic competition a market structure in which firms have many competitors, but each one sells a slightly different product.

duopoly two firms only in a market.

59

How the process of creative destruction is linked to technological change

The term **creative destruction** was first coined in 1942 by the Austrian economist Joseph Schumpeter to describe how capitalism, which dominates the economic system in which we live, evolves and renews itself over time. (Capitalism is the name given to the parts of the economy in which the means of production or capital are privately owned. In the UK, public limited companies or PLCs are the dominant form of capitalist business enterprise.)

In his famous book *Capitalism, Socialism, and Democracy*, Schumpeter wrote: 'The opening up of new markets, foreign or domestic…incessantly revolutionises the economic structure from within, incessantly destroying

CASE STUDY 2.6

Apple and creative destruction

On 1 April 1976, Apple Computer Inc. was incorporated by three 'techno-geeks', Steve Jobs, Steve Wozniak and Ron Wayne. Twenty-one years later in 1997, Steve Jobs, having left Apple following disputes about business strategy, rejoined the company and remained in charge until his death in 2011. (In 2007, Jobs had renamed the company Apple Inc. to reflect the fact that Apple had diversified away from computers into the iPod, the iPhone, the iPad and iTunes.)

Over this period, and particularly since 2001 when the iPod was first marketed, Apple had a crushing effect on specific competitors. According to Barry Ritholtz, writing in the *Washington Post* shortly after Job's death from pancreatic cancer, this was creative destruction writ large. Ritholtz argued that Jobs remade entire industries according to his unique vision. From music to film, mobile phones to media publishing and computing, Job's impact has been enormous.

Today, the triple threat of iPod/iPhone/iPad has left behind a wake of overwhelmed business models, confounded managements and bereft shareholders. The businesses which have been destroyed, or left as mere rumps of their former selves, include Hewett-Packard (HP), Nokia and Blackberry. According to Ritholtz, HP's printer business might still have some ink left in its cartridges, but its PC operations are hurting, gutted by sales of the iPad. HP's tablet entry, the TouchPad, was an unmitigated disaster, unable to compete with the iPad.

In 2007, the Finnish company Nokia totally dominated the mobile phone market. Many people thought that Nokia's lead was more or less insurmountable. But what has happened since is a reminder of just how quickly and completely the market power of a previously 'dominant' tech firm can disappear. However, following the introduction of Apple's iPhone in 2007, the previous market leaders Nokia and Blackberry began a rapid decline.

Under a headline 'Once-cool Blackberry fails to keep pace with rivals', China's *Morning Post* described how Blackberry, an early mover in the high-end mobile phone market, lost market share mainly to Apple's iPhone and to smartphones powered by Google's Android operating system. While Blackberry was considered perhaps the hippest if not the largest mobile phone maker several years ago, the company quickly lost momentum as it failed to keep pace with innovations from rivals. Gerry Purdy, an analyst at Compass Intelligence, said that 'The one gigantic issue facing Blackberry was the delay in getting into the smartphone market. And that was three years after the iPhone was released. So that's six years. The market was moving too fast.' Blackberry was too complacent, having become 'blinded' to competitive threats.

Even software giant Microsoft has suffered from Apple's innovation and marketing. Once Apple's main competitor in computer manufacturing and software, Microsoft has become vulnerable on multiple fronts. It has missed nearly every major trend in technology in recent years. Microsoft still has its cash cows Windows and its Office suite of products, but the company could lose out significantly to Apple in the next few years.

Follow-up questions

1 What is meant by 'creative destruction'?
2 This case study was written at the time of Steve Job's death in 2011. Find out what has happened to Nokia's and Blackberry's smartphone business in the years since then.

the old one, incessantly creating a new one. This process of Creative Destruction is the essential fact about capitalism.' Schumpeter also stated that 'The essential point to grasp is that in dealing with capitalism we are dealing with an evolutionary process.'

Creative destruction is strongly related to the processes through which technological change and innovation affect the ways in which businesses behave. It describes a process in which economic growth occurs in the economy as a result of new innovations creating more economic value than that being destroyed by the decline of the technologies the new innovations replace. Over time, societies that allow creative destruction to operate grow more productive and richer; their citizens benefit from new and better products and higher living standards. Creative destruction is central to the ways in which free market economies and mixed economies develop and change over time.

ACTIVITY

Cathode-ray television-tube manufacturers, video rental shops, high-street travel agents and bookshops have all in recent years been victims of creative destruction. Research how this has happened in one of these industries (or in an industry of your choice) and explain why you think consumers may or may not have benefited from the process.

SUMMARY

- The theory of the firm is the main part of business economics.
- A firm is a business enterprise which either produces or deals in and exchanges goods or services.
- The building blocks of the theory of the firm include production theory, cost theory and revenue theory.
- Production and cost theory divide into short-run and long-run theory.
- Production is a process or set of processes for converting inputs into outputs.
- The key concept in short-run production theory is the law of diminishing returns, also known as the law of diminishing marginal productivity.
- t is vital to understand, and to distinguish between, marginal returns, marginal costs and marginal revenue.
- The marginal return of labour is the extra output or the change in the quantity of total output resulting from the employment of one more worker, holding all the other factors of production fixed.
- Marginal cost is the extra cost a firm incurs when it produces one extra unit of output.
- Marginal revenue is the extra sales revenue a firm receives when it sells one extra unit of output.
- In the short run, the marginal cost curve and the average variable cost curve are derived from the law of diminishing marginal (and average) returns.
- Assuming that the variable factors of production experience diminishing returns, the average variable cost (AVC) and the short-run average total cost ($SRATC$) curves are U-shaped.
- The key concept in long-run production theory is returns to scale.

61

- Increasing returns to scale, constant returns to scale and decreasing returns to scale are all possible.
- Increasing returns to scale are likely to lead to economies of scale, which are defined as falling long-run average costs.
- Decreasing returns to scale are likely to lead to diseconomies of scale, which are defined as rising long-run average costs.
- The long-run average cost (LRAC) curve may be U-shaped, but other shapes are possible.
- Minimum efficient scale (MES) is the lowest output at which the firm is able produce at the minimum achievable LRAC.
- Profit is total sales revenue minus total costs of production.
- Internal economies and diseconomies of scale should not be confused with external economies and diseconomies of scale.
- A firm's sales revenue, which must not be confused with returns to factors of production, is influenced by the market structure in which the firm sells its output. A key difference is that returns relate to physical units of output whereas revenue is measured in terms of money.
- Perfect competition, monopoly, monopolistic competition and oligopoly are the four market structures you need to know, and you must also be aware of the meaning of imperfect competition.
- A perfectly competitive firm is a price-taker, but a monopoly is a price-maker or quantity-setter.
- Economists usually assume that maximising profit is a firm's main business objective.
- Profit is total sales revenue minus total costs of production.
- Normal profit is just sufficient to keep incumbent firms in the market but is insufficient to attract new firms into the market.
- Normal profit is treated as a cost of production and is included in a firm's cost curves.
- Abnormal, or supernormal, profit is any profit over and above normal profit.
- The ways in which firms operate are affected by technological change, which encompasses the processes of invention, innovation and diffusion of technology in the economy.
- In the long run, capitalism develops through a process known as creative destruction, through new technologies and innovations replacing older technologies and innovations.

Questions

1 Explain the difference between the law of diminishing returns and decreasing returns to scale. How do these affect a firm's cost curves in both the short run and the long run?

2 What is the relationship between returns to scale and economies and diseconomies of scale?

3 Explain why the average and marginal revenue curves of a monopoly slope downward, while those of a perfectly competitive firm are horizontal.

4 Explain the mathematical relationships between the average and marginal values of an economic variable.

5 Evaluate the view that a monopoly can simultaneously increase both the price of the good it produces and the quantity of the good it sells.

6 With two examples of each concept, explain the difference between invention and innovation.